UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

STATISTICAL METHODS FOR MODELING AND *NOWCASTING* THE IMPACTS OF INFLUENZA EPIDEMICS

Baltazar Emanuel Guerreiro Nunes Bravo Nunes

DOUTORAMENTO EM ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL
(Probabilidades e Estatística)

2011

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

STATISTICAL METHODS FOR MODELING AND *NOWCASTING* THE IMPACTS OF INFLUENZA EPIDEMICS

Baltazar Emanuel Guerreiro Nunes Bravo Nunes

Tese orientada pela Professor Doutora Maria Lucília Salema Carvalho e pela Professora Doutora Isabel Cristina Maciel Natário

DOUTORAMENTO EM ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL
(Probabilidades e Estatística)

2011

Title: Statistical methods for modeling and *nowcasting* the impacts of influenza epidemics.

Author: Baltazar Emanuel Guerreiro Nunes Bravo Nunes

Supervisor: Professor Doctor Maria Lucília Salema Carvalho and Professor Doctor Isabel Cristina Maciel Natário

Abstract: Influenza is an acute respiratory infection responsible for epidemics with high impact on human health. Several statistical methods have been applied to data collected from influenza surveillance systems (ISS) to assess the epidemic burden and early detect it. Given the ISS reporting delays, models have recently been developed to correct them by predicting the present situation (*nowcasting*) using the incomplete information collected. Thus, three objectives were defined.

Review and classify the methods that use interrupted mortality time series to estimate influenza excess deaths. They were classified according to the model used to fit the time series and obtain a baseline; the influenza epidemic period estimator and the procedure used to fit the model (iterative or non iterative). This generalization led to the development of user friendly `R`-package, `flubase`, implementing all these models.

Estimate influenza excess deaths in Portugal between 1980 and 2004. The seasonal excess deaths average by all causes was 2,475, of those 90% occurred in the elderly. These results suggest a similar influenza epidemics profile between Portugal and other countries in the Northern Hemisphere, and represent the first reference to contextualize future epidemics severity and design public health measures.

Develop a model to nowcast the influenza epidemic evolution in a weekly basis. A non homogenous hidden Markov model (HMM) was developed to nowcast the current week influenza-like illness (ILI) incidence rate and the probability that the influenza activity is epidemic using as covariates an early estimate of ILI rate and the number

of ILI cases tested positive in the previous week. Bayesian inference was used to estimate the model parameters and nowcasted quantities. The results obtained by application to the Portuguese ISS data, demonstrated the additional value of using a non homogenous HMM instead of an homogenous since it improves the ISS timeliness in 2 weeks.

Key words: influenza, epidemics, baseline, excess deaths, cyclical regression, Autoregressive Moving Average (ARIMA) models, hidden Markov models (HMM), non homogenous HMM, bayesian models, Markov Chain Monte Carlo (MCMC), nowcasting, surveillance.

Título: Statistical methods for modeling and *nowcasting* the impacts of influenza epidemics.

Autor: Baltazar Emanuel Guerreiro Nunes Bravo Nunes

Orientadores: Professora Doutora Maria Lucília Salema Carvalho e Professora Doutora Isabel Cristina Maciel Natário

Resumo:

A gripe é uma doença respiratória aguda que no hemisfério norte, durante o Outono e Inverno, é responsável por epidemias com considerável impacto nas populações humanas, traduzindo-se muitas vezes em excessos de mortalidade, hospitalizações e necessidades de cuidados de saúde.

Neste contexto, têm sido implementados vários sistemas de vigilância epidemiológica da gripe (SVG) a nível nacional e internacional com o objectivo de fornecer às autoridades de saúde informação para a elaboração de avaliações de risco actualizadas que permitam uma correcta implementação de medidas de controlo e mitigação das epidemias e suas consequências.

Com o objectivo de medir o efeito das epidemias em termos de excessos de mortalidade e de detectar de forma precoce o seu início, diversos métodos estatísticos têm sido propostos e aplicados aos dados colhidos por estes SVG. Em relação a este último objectivo, e dado que muitos dos SVG apresentam importantes demoras no processo de recolha, tratamento e análise dos dados, com consequentes atrasos na detecção das epidemias, recentemente têm-se desenvolvido modelos estatísticos que procuram corrigir estas faltas de informação. Os modelos propostos procuram então prever a situação epidémica actual - *nowcasting* - usando a informação incompleta colhida até ao momento pelo SVG.

Neste enquadramento os objectivos desta tese foram: unificar numa única classe os métodos estatísticos para estimar os excessos de mortalidade atribuíveis à gripe,

que são caracterizados por usarem séries temporais de mortalidade interrompidas; estimar os excessos de mortalidade atribuíveis à gripe durante o período de 1980 a 2004 e contextualizá-los na literatura ciêntifica internacional; e desenvolver modelos para prever a presente situação epidémica da gripe (nowcast) no contexto dos sistemas de vigilância epidemiólogica.

Os principais métodos estatísticos, que recorrem a séries temporais da mortalidade interrompidas para estimar os excessos de óbitos associados à gripe foram revistos de forma exaustiva. O objectivo foi identificar não só as suas características comuns mas também os factores que os diferenciam. Desta análise resultou uma unificação dos métodos que se caracterizou pela sua classificação de acordo com os seguintes parâmetros: o tipo de modelo usado para ser ajustado à série temporal interrompida e estimar a linha de base (regressão cíclica ou ARIMA), o período temporal escolhido *a priori* usado para estimar o período epidémico e o procedimento para ajustar o modelo à série temporal (iterativo ou não iterativo). Esta gereralização e formalização levou naturalmente à construção de um conjunto de rotinas de `R` de fácil utilização, o pacote `flubase`, que pode ser descarregado de `http://cran.r-project.org/web/packages/flubase/index.html` e onde estão implementados todos os métodos descritos. O pacote de rotinas desenvolvido representa também uma importante ferramenta para a avaliação da sensibilidade dos excessos de óbitos obtidos face à variação do tipo de método usado, pois permite obter de forma prática e rápida estas estimativas para diferentes combinações dos parâmetros. Os vários métodos identificados foram ainda aplicadas a 20 anos de mortalidade por Pneumonia e Gripe em Portugal, demonstrando que, neste caso, o parâmetro que maior impacto teve nas estimativas dos excessos de óbitos foi o tipo de período escolhido para estimar o período epidémico.

Com base nos resultados obtidos no estudo anterior seleccionou-se o método estatístico considerado mais adequado à estimação retrospectiva dos excessos de mortalidade associados à gripe em Portugal. Mais específicamente foi aplicado às séries de mortalidade estudadas o modelo ARIMA com ajustamento não iterativo onde os períodos epidémicos foram estimados com base na mortalidade específica por gripe. Os resultados obtidos da aplicação do método a 7 causas de morte diferentes para 8 grupos etários, permitiu: estimar em 2.475 a média sazonal de óbitos associados às epidemias de gripe que ocorreram no período de 1980 a 2004, valor este que corresponde a uma taxa média bruta por época de 24,7 por 100.000 habitantes; verificar

que em 5 das 24 épocas não ocorreram excessos de óbitos associados à gripe e que o máximo estimado foi de 8.514 óbitos na época 1998-1999. Um outro resultado importante foi que em média, os excessos estimados no grupo etário ≥ 65, representaram cerca de 90% do total dos excessos. Todos os resultados obtidos sugerem ainda que as epidemias de gripe ocorridas neste período em Portugal tiveram, em termos gerais, um perfil semelhante ao descrito noutros países com clima temperado do Hemisfério Norte. Adicionalmente, poderemos ainda afirmar que as estimativas obtidas neste estudo, representam um passo importante para estabelecer referências para avaliar o impacto de futuras epidemias de gripe e também para delinear medidas de saúde pública racionais para mitigar o seu efeito.

O objectivo de reduzir a demora dos SVG na detecção do início do período epidémico, foi atingido com a apresentação de um modelo que permite prever a taxa de incidência de síndroma gripal (SG), assim como o estado de actividade gripal (epidémico ou não epidémico) da própria semana. Este modelo foi escolhido na família de modelos de cadeias de Markov escondidas (HMM), porque aplicações anteriores, no contexto da detecção de epidemias, demonstraram algum sucesso e principalmente porque estes modelos permitem a previsão simultânea de duas medidas de grande interesse para este trabalho - a probabilidade de se estar no período epidémico e a taxa de incidência de SG. Nestes modelos as probabilidades de transição entre estados podem ser assumidas como contantes ou variantes no tempo, correspondendo respectivamente ao modelo homogéneo e não homogéneo. Assim, elegeu-se naturalmente o modelo não homogéneo para atingir o objectivo definido, dado que tem a vantagem de permitir a inclusão de covariáveis com informação precoce sobre a evolução da epidemia que permitem, ao mesmo tempo, modelar a variável resposta, taxa de incidência de SG, mas também as probabilidades de transição entre o estado epidémico e o não epidémica e vice versa.

As covariáveis escolhidas foram uma estimativa precoce da taxa de incidência de SG calculada à sexta-feira da própria semana e o número de casos de SG com resultado laboratorial positivo para gripe na semana anterior. As estimativas dos parâmetros dos modelos assim como a taxa de incidência e a probabilidade de estar no estado epidémico foram obtidas por métodos de inferência bayesiana. Os resultados obtidos pela aplicação dos modelos propostos à informação recolhida pelo SVG Português, demonstraram a vantagem de usar um modelo de cadeias de Markov escondidas não homogéneo em comparação com um modelo homogéneo. Concretamente foi possível

monstrar que, no caso deste SVG, o recurso a um HMM não homogéneo reduz o atraso na detecção do início do período epidémico em duas semanas.

Palavras chave: influenza, epedimias, linha de base, excessos de mortalidade, regressão cíclica, Modelos Autoregressivos Intregados de Médias Móveis (ARIMA), Modelo de cadeias de Markov escondidas (HMM), HMM não homogéneos, modelos bayesianos, Markov Chain Monte Carlo (MCMC), nowcasting, vigilância.

À minha mãe

Acknowledgements

It is pleasure to thank to those who made this thesis possible:

- Professor Doctor Maria Lucília Carvalho, my supervisor, for being there in all the steps of this work, for giving me all the support, clear guidance and most of all for the enthusiasm and motivation I have needed. I also acknowledge all the teachings and interest, in probabilities and statistics, that Professor Lucília have gave me during my formation as a statistician during my degree and Master degree courses.

- Professor Doctor Isabel Natário, my co-supervisor, for the systematic support and incentive. And specially for maintaining me on the track of the discipline and methodological rigor, and to insist in formal demonstration before data analysis. That was clearly a very important lesson for me.

- Dr. José Marinho Falcão, my boss during 12 of the 14 years of my career. I owe him much of the experience and knowledged in epidemiology and public health research. With this shoulder by shoulder experience, I had the privilege of learning to focus on the objectives and the relevance of the results and mainly to be transparent and clear in all the research process. Like he would say citing George Orwell "...telling the truth is a revolutionary act."

- Dr. Carlos Matias Dias, the head of the Department of Epidemiology of the Instituto Nacional de Saúde Dr. Ricardo Jorge (INSA), for giving me all the support I needed, and setting as a priority for me in the department the achievement of the doctor degree. Without his aid I would not have the time neither the space to conclude this thesis. I also thank him for the lessons on how to write a PhD thesis in an home with small children in need for attention and willing to help in everything.

- Doctor Cecile Viboud and her colleagues at the Fogarty International Centre of the National Health Institutes in Bethesda, MA, USA, for receiving me during 3 months in order to develop the content of Chapter 3 of the thesis. During this period and afterward I had the privilege of working with a top research group in the context of the Multinational Influenza Mortality Study, funded by the International Influenza Unit, Office of Global Health Affairs, Department of Health and Human Services, USA, that clearly contributed to the enrichment of my scientific abilities.

- Professor Doctor José Pereira Miguel, President of the Instituto Nacional de Saúde Dr. Ricardo Jorge, and the management board, for giving me the possibility to work in this thesis in leave of absence for two periods of 3 months.

- Professor Doctor Antónia Turkman, as coordinator of the Centre of Statistics and Applications of the University of Lisbon, for the financial support that CEAUL gave me to present the work developed in this thesis in two international conferences.

- "Médicos-Sentinela" The Portuguese general practitioners network for voluntarily providing weekly data for the influenza surveillance during more than 20 years.

- A special thanks go also to Zilda Pimenta, Inês Batista and Dr. Isabel Falcão, colleagues of the Department of Epidemiology, for weekly maintaining the surveillance system working like clock and most of all for being there always when I needed.

- Dr. Raquel Guiomar, coordinator of the National Influenza Reference Laboratory of INSA for the support and for providing the data from the laboratorial component of the Portuguese influenza surveillance system.

- Engineer Ausenda Machado for all the help and advise in the discussion of the Chapter 3 results.

- All my colleagues of the Department of Epidemiology of INSA for being there all the time and being my solid ground during this path. I also thank them for sustaining the department tasks during my physical and mental absences.

- Doctor Helena Rebelo-de-Andrade for the advises in the development of Chapter 3 and also for introducing me in the area of influenza research.

- FCT - Fundação para a Ciência e a Tecnologia (National Funds) - in the scope of project PEst-OE/MAT/UI0006/2011 for having partially funded this work.

- Luso-American Foundation for financing my three months visit to the National Health Institutes, Bethesda, MA, USA with the project 1-13/08.

- Baltazar Bravo Nunes, my father, for once more contributing with funds for the PhD tuition fees as *pro bono*. Even at this point of my adult live he continues to be my safe port.

- To finalize, the most important ones, Inês and Margarida, for resenting at each moment my priorities and for filling my life with sense and love.

Notation

The following conventions are generally followed:

- Random variables by upper-case letters and observed values of these by the corresponding lower-case letters.
- Greek letters are used to denote parameters.
- Bold symbols correspond to vector and matrix notation.

List of Abbreviations

- ARIMA - Autoregressive Integrated Moving Average;
- ECDC - European Center for Prevention and Disease Control;
- GP - General Practitioner;
- HMM - Hidden Markov Models;
- ICD - International Classification of Diseases
- ILI - Influenza-like Illness;
- ISS - Influenza Surveillance System;
- MCMC - Markov chain Monte Carlo;
- P&I - Pneumonia and Influenza;
- RSV - respiratory syncytial virus;
- US CDC - United States Center for Disease Control and Prevention;
- WHO - World Health Organization.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Influenza is an acute respiratory infection that, in the temperate climates, during Autumn and Winter, is responsible for epidemics of considerable dimension, with attack rates that vary between 5 and 10% of the general population and with duration of 2 to 26 weeks [1, 2].

During these epidemic periods, influenza is associated with an increase of morbidity and mortality from all causes, mainly in the individuals with 65 or more years of age. Nevertheless the burden of influenza can also be high in the younger age groups, usually during pandemics, like 1918-19, 1957, 1967 and 2009-10 [3, 4, 5].

In Europe it is acknowledged that seasonal influenza epidemics are responsible for an average of 40,000 deaths per season, an important increase in the need for health services capacities and also for a big impact in the labor force given the considerable absenteeism they originate [6]. Given the substantial impact of the influenza epidemics in the human populations, public health surveillance systems were established since 1952, when the World Health Organization (WHO) launched the WHO Global Influenza Surveillance Network. This network is based on the National Influenza Centers that collect and send biological samples from patients with influenza-like illness (ILI) to WHO Collaborative Centers for antigenic and genetic analysis. The main objective of this network is to recommend the content of the influenza vaccine and serve as a global alert mechanism for an influenza virus with pandemic potential [7].

Further more, to monitor the impact of influenza epidemics in the mortality and morbidity of the populations other systems were also implemented. Examples are: the *122 Cities Mortality Reporting System*, managed by the United States Center

for Disease Control and Prevention (US CDC), that collects weekly the number of deaths due to Pneumonia and Influenza (P&I) by age group in 122 cities of the US [8]; the Euro-MOMO project that collects the weekly mortality from all causes in a network of 13 European countries [9]; the European Influenza Surveillance Network that is based on sentinel networks of general practitioners (GP) that report on a weekly basis the ILI incidence rate [6]. All these surveillance systems have the objective of supplying to the health authorities information to perform up-to-date risk assessments for public health action, i.e. implement measures to control and mitigate the epidemics impact. Generally, information provided by these systems helps the health authorities in following the influenza epidemic evolution, week by week, in terms of the epidemic onset, peak and end, and also in terms of its impact and severity, measured as excess of medical consultations, hospitalizations and mortality.

For these purposes, several statistical methods have been proposed and applied to the data collected by these surveillance systems. The reason for the first studies was the 1918-19 influenza pandemic. To access the impact of this event, Collins SD (1957) [10] suggested an ecological method that estimates the monthly expected mortality rate in the absence of the influenza epidemic and subtracts the observed mortality rates from the expected. The sum of these excesses during the epidemic period was considered as the excess mortality attributable to the pandemic.

In fact, this rational of defining *baselines* that describe an indicator behavior in the absence of public health threats has been the basis for the majority of the statistical methods that were presented, from then to nowadays, to identify the start of a public health event and to estimate its impact.

Another important feature of the public surveillance system, operating for several years, is the richness of information it contains, which could allow short term forecasting of indicators under analysis. For some situations, this short term forecasting is indeed a *nowcasting*, given that surveillance systems usually report information with a delay of one or more weeks, and a one or two weeks forecast is, in practice, a prediction of the current situation, i.e. a nowcasting. The application of statistical methods to predict the present situation or the near future, can greatly contribute to reduce the surveillance system timeliness and enhance the up-to-date epidemic risk assessment.

On this subject some works have also been presented with relative success [11, 12], nevertheless none has shown to be sufficiently practical to be implemented week by

week as a new outbreak indicator of the surveillance system.

In this context, the main research objectives of this thesis are:

1. To unify in a single class the statistical methods characterized by using interrupted mortality time series to estimated excess deaths attributable to influenza epidemics, in order to describe and compare their applicability and results;

2. To estimate the excess mortality associated with the influenza epidemics occurred in Portugal in the period from 1980 to 2004 and compare the results with those from other locations;

3. To develop a statistical model to nowcast the influenza epidemic evolution.

In general, for the first objective, the research focused on the group of methods characterized by not considering influenza activity covariates to model the mortality baseline and by excluding from the model fitting process all the parts of the mortality time series where there was evidence of influenza epidemics occurrence. So, to describe this group a comprehensive review of the main proposed methods [13, 14, 15, 16, 17, 18, 19] was carried out with the aim of finding not only their unifying characteristics but also their differences. The identified features were then used to set a general framework that encompasses them all. Finally, to compare the methods and the impact of each feature on the estimates of the excess deaths associated with influenza epidemics, the different methods were applied to the time series of deaths by P&I in Portugal from 1980 to 2004 [20].

To accomplish the second goal the method that in [20] was shown to be more appropriate to retrospectively estimate the excess deaths attributable to influenza epidemics was elected to be applied to the Portuguese mortality data in the period of 1980 to 2004. So, given that the last study on the burden of influenza epidemics in Portugal [21] was focussed only on excesses of all causes and P&I deaths for all ages and for the elderly ($\geq$ 65 years) in the period from 1990 to 1998, this new work [22] has analyzed the mortality time series for 7 causes of death (all causes, cerebrovascular disease, ischemic heart disease, diseases of the respiratory system, chronic respiratory diseases, and pneumonia and influenza) and 8 age groups (0-4, 5-54, 55-64, 65-69, 70-74, 75-79, 80-84 and $\geq$ 85).

Finally, to achieve the last research objective, to nowcast the influenza epidemics evolution, in a public health surveillance system setting, a statistical model, within

the family of the hidden Markov models (HMM), was developed. This class of models was chosen not only because some success in outbreak detection problems has already been accomplished with this framework [23, 24, 25], but also because it was further noticed that it can enable the nowcast of two important measures, simultaneous: the probability of being in the epidemic state and the ILI incidence rate. Within the HMM family the state transition probabilities can be assumed constant in time or time-variant, respectively corresponding to a homogenous or non homogenous HMM. For this work, the non homogenous model was the elected one, because it has the advantage of allowing the inclusion of time-variant covariates with early information on the epidemic evolution to model not only the response variable, the ILI weekly rate, but also the state transition probabilities from the non epidemic to the epidemic state and *vice-versa*. To our knowledge, this work [26] represents the first attempt to use non homogenous HMMs in a disease surveillance problem with the objective of early detect an outbreak and nowcast its evolution.

The thesis is organized as follows. Chapter 2 focuses on the methods used to estimate excess deaths attributable to influenza epidemics. Hence, after an introduction to these methods (Section 2.1) and the presentation of some essential concepts (Section 2.2) a description of the most relevant methods is made in Section 2.3. A general framework proposed with the objective of finding a common ground between the methods, to describe and compare them, is then given in Section 2.4. The comparison between the methods is exemplified by the application of all class methods to the time series of deaths due to P&I in Portugal from 1980 to 2004 (Section 2.5). In Section 2.6 a set of user friendly `R`-routines, package `flubase` [27], which implement all methods is briefly presented. The main results are then discussed in Section 2.7.

The estimates of excess mortality rates associated with influenza virus circulation in Portugal, during the period of 1980 to 2004, is presented in Chapter 3. An introduction to the impact of influenza epidemics in terms of excess mortality is given in Section 3.1. The data description is presented in Section 3.2 and the method used to estimate the epidemic periods and the excess deaths (selected from the framework presented in Chapter 2) are presented in Sections 3.3 and 3.4. Section 3.5 describes the main results and the specificity analysis applied to evaluate the method robustness. Finally, Section 3.6 frames the obtained estimates in the published literature and discusses the main differences and similarities between countries, geographical regions and population characteristics.

In Chapter 4 the development of a non homogenous HMM to nowcast the influenza activity in the context of a public health surveillance system is presented. Section 4.1 introduces the motivation, the question of timeliness of a public health surveillance system and the need to have predictions of the current situation. An overview of the HMMs and their application to the influenza surveillance problem is then given in Section 4.2, along with the formalization of the non homogenous HMM. At Section 4.3, the data used for the application example is introduced, the Portuguese influenza surveillance system (ISS) from week 40/2008 to week 16/2011. In Sections 4.4 and 4.5 the specific models proposed are described along with the bayesian approach for the model parameters estimation and for nowcasting the current ILI rate and influenza activity state. Section 4.6 details the results, both the application of the models to the entire data set as well as the real-time nowcast of the 2010-11 influenza season. In section 4.7 the model and results are discussed.

Finally Chapter 5 presents the main conclusions about the research objectives and suggestions for future developments.

Chapter 2

Time series methods for obtaining excess mortality attributable to influenza epidemics[1]

2.1 Introduction

In the Northern Hemisphere countries, during influenza epidemic periods, a rise in mortality from all causes is usually observed, mainly in the elderly population (aged 65 years or more) [28]. This increase can be associated with influenza epidemics since the influenza infection might cause complications that can lead to the hospitalization and/or death of the infected individual [36, 37]. In this context, and from a public health point of view, the quantification of the influenza epidemics impact on the population and its description in terms of the dominant virus strain and level of vaccine coverage is of the utmost importance. The measurement of influenza impact in terms of deaths hospitalizations is never accessed by the number of deaths with influenza as the main cause in national mortality registries because this value is usually very low, even during the most severe epidemics. This is mainly due to the difficulty in estab-

[1]This Chapter is based on the paper Nunes B, Natário I, Carvalho ML. Time series methods for obtaining excess mortality attributable to influenza epidemics. Statistical Methods in Medical Research. 2011; 20(4):331-346. Epub 2010 March 8.

lishing a connection between a complication (pneumonia or other respiratory diseases, circulatory system diseases, etc) and a previous or current influenza infection, due to the lack of a laboratory confirmed diagnosis. As a consequence, the use of official death registries, with influenza as cause to measure the influenza epidemic impact would underestimate its real effect [28]. This has led researchers to look for reliable methods to estimate influenza-associated deaths, using as a starting point a mortality time series and, when available, additional information from influenza epidemiologic surveillance systems on the seasonal epidemic characteristics. Generally, the methods used to estimate the excess deaths attributable to the influenza epidemics follow three steps:

1. Obtaining a baseline of the number of deaths, by a certain time unit, in the absence of influenza epidemics;

2. Using the baseline to identify the periods where there is evidence of an excess of deaths attributable to influenza epidemics;

3. Subtracting this baseline from the observed number of deaths, during these periods.

In this sense, the observed excess of deaths above the baseline, when associated with influenza epidemic periods, could, in the absence of other explainable events, be attributed to an influenza epidemic. The state-of-the-art of the methods to estimate the excess deaths attributable to influenza epidemics offers a large variety of different alternatives, all applicable to identical situations and aiming essentially the same purpose. These methods can be classified into two general methodological approaches and, within those, they vary in a considerable number of aspects.

In the first group, the methods are based on statistical models that include influenza activity indicators as explanatory covariates. The pioneer ones [29, 30, 31] are multiple regression models including a polynomial function of time for trend, dummy variables for month effect to cope with seasonal variation, the monthly mean of weekly minimum temperature and the monthly ILI incidence rate as an influenza activity indicator. Later, [32, 33, 34] propose a Poisson regression model including, in general, the same type of variables, except [33] that uses one influenza activity indicator by each sub-type of influenza virus, A(H3), A(H1) and B, plus an indicator of the respiratory syncytial virus activity, all considered as proportion of isolates by week. Thus,

this type of method enable the estimation of influenza burden by sub-type taking into account the possible simultaneous effects of other factors in mortality, like the effect of climate and/or other respiratory infections. These methods are very exigent in terms of external data needed and are also dependent on the accuracy of the influenza activity indicators used that are, in general, based on sentinel surveillance systems, which are known to be sometimes influenced by external factors like holiday periods (e.g Christmas and New Years Eve) [35].

The methods in the second group are characterized by not considering covariates and also by excluding from the estimating process all the parts of the mortality time series where there is evidence of influenza epidemics occurrence. This Chapter will be focused on this second group of methods. A close analysis of this group shows several differences among them, essentially on the type of statistical model employed (cyclical regression [14, 15, 16, 17, 18] or Autoregressive Integrated Moving Average (ARIMA) models [13, 19]), on the method used to build the baseline (non iterative [17, 16, 18] or iterative [14, 15, 13, 19]) and on the choice of the periods to be excluded from the mortality time series (epidemic periods defined using ILI surveillance systems [14, 17, 13, 19] or fixed periods, like December to April [15, 16, 18]). All these differences can lead to unequal influenza-associated deaths estimates. Differences between reported estimates have been identified leading specialists into a profound discussion without a final agreed conclusion [38, 39].

Here we were able to unify these methods in a single class, in such a way that it allows the description and comparison of their applicability and results. This proved to be an important step in the conceptualization of the statistical methods used to estimate influenza-associated deaths, clarifying all the steps performed and options taken to compute the desired estimates. This unification was also the basis to build an `R`-package, the `flubase`, that easily estimates the influenza-associated deaths by any of the methods in the class. This platform is quite user friendly even for those less familiarized with the theoretical statistical developments that have led to the results. The application of this tool could also empower other researchers to critically analyze the differences and similarities between the estimates obtained with a variety of method choices, allowing in this way a more comprehensive analysis of their data.

2.2 Some essential concepts

2.2.1 Influenza epidemic period, influenza season and flu-year

An *influenza epidemic* is defined as the occurrence, in a specific population, of a number of cases of influenza above what is usually expected, during a certain period of time, referred as the *epidemic period*. Usually the epidemic periods are unknown and therefore must be estimated.

The annual fixed period of time during which the influenza epidemics might occur, starting sooner or later, with larger or smaller duration, is named *influenza season*. In Portugal as in other northern hemisphere countries, this period starts in October of each year ending in May of the next calendar year. This is the period when the ISS are more active, as the occurrence of an influenza epidemic outside this period has an almost null probability. Taking into account the beginning and ending of the influenza season, a *flu-year* a is defined as the 52 (or 53 when the first calendar year of the *flu-year* is bissextile) weeks that start at week 27 of any calendar year n and ends at week 26 of the calendar year $n + 1$.

Let E_a denote the estimate of an influenza epidemic period occurred during flu year a (Figure 2.1). The choice of these periods is greatly dependent on the level of information one has on the occurrence of influenza cases and on the temporal evolution of the influenza incidence rate in the population. In fact, to obtain a correct diagnosis of influenza, a confirmation of the influenza virus presence is necessary, procedure that is not usually carried out. In the majority of the situations only the clinical diagnosis are obtained, without the laboratory confirmation, and if this situation occurs the case can only be classified as ILI.

Therefore, the information that is usually available consists on the temporal evolution of the ILI incidence rates, complemented by information on the influenza virus circulation among the population. In the majority of the developed countries this information is collected by surveillance systems specifically designed for the effect, that are based on a sample of individuals set under surveillance.

When this information does not exist, or it is not available, some authors [16, 18] have set E_a as the fixed time period, enclosed in the influenza season period, that goes from December to April of the next calendar year. Other solution is to use the time series of mortality specific by influenza (ICD 9th Revision:487; ICD 10th Revision:J9-J11) and define E_a as the periods where the mortality by influenza rises

above the expected. In principle, given the under registration of deaths with influenza as a cause, this last option should be a less sensible but more specific method, for the reasons we have mentioned earlier.

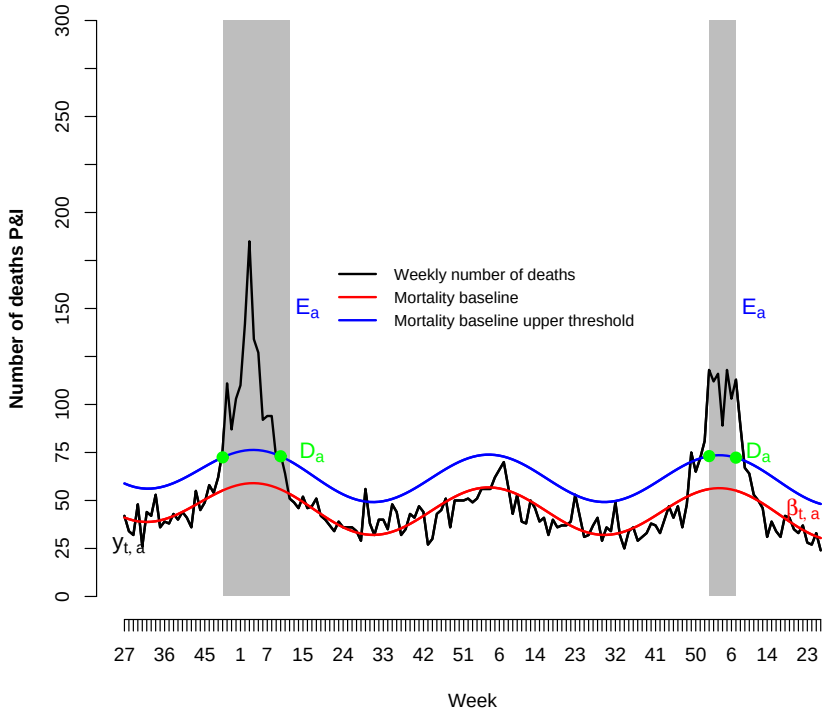

Figure 2.1: Exemplification of the basic concepts.

2.2.2 Time series of the weekly number of deaths

Consider $y_{t,a}$ to be the time series of the number of deaths observed in week t, $t = 1(27), ..., 52(26)$[2] of flu-year a, $a = 1, ..., A$, where A represents the number of flu-years in study. This time series is the main object of analysis, since the major goal is to estimate from it the excess number of deaths attributable to influenza epidemics. Usually the most used time series is the weekly (or monthly) number (or rate) of deaths by all causes or P&I.

[2]The values in brackets represent year calendar week numbers

2.2.3 Periods with excess deaths attributable to influenza epidemic

Define D_a as the period of weeks where **an excess of deaths** in $y_{t,a}$ is attributed to an influenza epidemic, in flu-year a. This period, included in the E_a period, is defined by an observed increase in $y_{t,a}$, above the expected in the absence of the effect of an influenza epidemics (Figure 2.1). Additionally, during this period, there must be no other events that can be the cause of the observed excess deaths.

2.2.4 Mortality baseline in the absence of the influenza epidemics effect

Excluding from the time series the parts where there is some evidence of an influenza epidemic occurrence, one obtain the following *interrupted time series*, denoted hereafter by $y^{\star}_{t,a} = \{y_{t,a} : (t,a) \notin E_a\}$.

Let $\beta_{t,a}$ be the baseline (Figure 2.1) resulting from fitting a statistical model to the time series $y^{\star}_{t,a}$ or, as some authors have considered, to $x_{t,a}$, the weekly number of deaths in the absence of the influenza epidemics, defined as follows:

$$x_{t,a} = \begin{cases} y_{t,a}, & (t,a) \notin D_a; \\ \tilde{y}_{t,a}, & (t,a) \in D_a \end{cases} \tag{2.1}$$

where $\tilde{y}_{t,a}$ represents some preliminary estimate of the weekly number of deaths in the absence of the effect of an influenza epidemics for the week t of the flu-year a.

2.3 Description of methods in study

The studied methods are all characterized by obtaining a mortality baseline in the absence of influenza epidemics effects using an interrupted mortality time series. Generally these methods fit a statistical model to $y^{\star}_{t,a}$ to obtain a baseline $\beta_{t,a}$ that is used to identify the periods with excess deaths attributable to an influenza epidemic, D_a.

To be able to jointly describe all these procedures one has to identify the unifying characteristics and also their differences in order to summarize them in a few classes. Three sources of dissimilarity were found.

1. **Statistical model used to fit the interrupted time series**: There are mainly two types of models used in the literature:

(a) multiple linear regression models [14, 15, 16, 17, 18], using a polynomial component to explain the series trend and a sinusoidal component that captures the seasonality observed - cyclical regression. Generally these models are given by:

$$y_s = \alpha + \sum_{i=1}^{m} a_i s^i + \sum_{j=1}^{l} b_{1,j} \sin \frac{j2\pi s}{52} + b_{2,j} \cos \frac{j2\pi s}{52} + \varepsilon_s$$

where $a_i,\ i = 1, ..., m$, are the parameters of the order m polynomial function used to explain the trend, $b_{1,j}$ and $b_{2,j}$ are the parameters of the sinusoidal function with periods $52/j,\ j = 1, ..., l$, used to explain the eventual seasonality and $\varepsilon_s \sim N(0, \sigma^2)$ with $s = t + (a-1)52,\ t = 1, ..., 52$ and $a = 1, ..., A$.

(b) seasonal ARIMA[40] only applied in [13, 19] to this problem.

2. **Choice of the $\mathbf{E_a}$ periods**: In some of the reviewed papers this period was the epidemic period defined (estimated) by the operating ISS, using data on clinical diagnosis of ILI and viral strains isolates [14, 17, 13]. In this case the a priori chosen E_a periods are different from flu-year to flu-year.

   Other authors [15, 16, 18] defined E_a as a fixed set of weeks (December to April), in each flu-year, where the occurrence of an influenza epidemic with effects on mortality is more likely. This period is always included in the influenza season.

3. **Procedure used to fit the statistical model and to identify the $\mathbf{D_a}$ periods**

   (a) Non iterative (Figure 2.2): the model is fitted to all points of the interrupted time series $y^{\star}_{t,a}$ [16, 17, 18] at once. Here the baseline $\beta_{t,a}$ corresponds to the estimated values given by the model for each week t. In [17] the D_a periods are defined as the set of weeks, contained in the E_a periods, that initiate with two consecutive weeks with a number of deaths above the upper 95% confidence limit of the baseline and end with two consecutive weeks with a number of deaths bellow the same upper limit. Note that [16, 18] have defined the D_a periods applying the previous method only to the mortality time series specific for influenza.

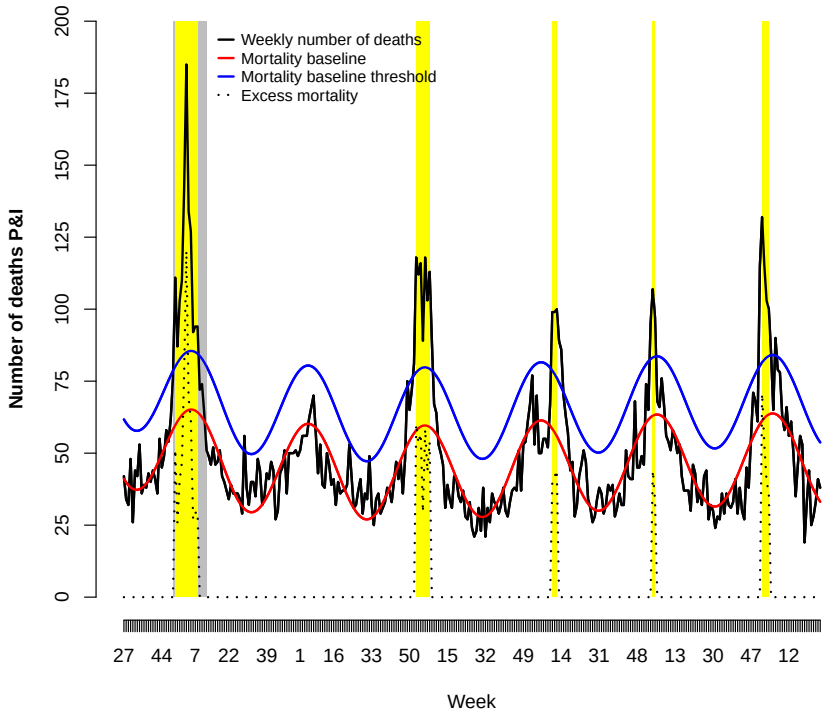

Figure 2.2: Non iterative procedure used to fit the statistical model and to identify the $\mathbf{D_a}$ periods. Grey boxes represent the E_a periods and yellow boxes represent the D_a.

(b) Iterative (Figure 2.3): generally, these methods consist in forecasting a baseline for each flu-year i using a statistical model fitted to $x_{t,a}$ for a training set of T previous flu-years. This training set can have a fixed dimension T (equal for all iterations, e.g. 5 years) [14, 15], or be given by all previous years of flu-year i of that iteration [13]. In the iterative methods, D_i is identified in each iteration i as the period of weeks contained in the correspondent E_i period that initiates with two consecutive weeks with a number of deaths above the upper 95% confidence limit of the forecasted baseline and terminate with two consecutive weeks with number of deaths bellow the same upper limit. After the D_i identification the series $x_{t,a}$ is updated. Here [13] update $x_{t,a}$ by substituting its values in the D_i periods

by the values of the forecasted baseline, obtaining a preliminary estimate of the mortality time series in the absence of influenza epidemic $\tilde{y}_{t,a}$ during those periods. Other authors simply use an interrupted time series where the values in the D_i periods are excluded from the original time series.

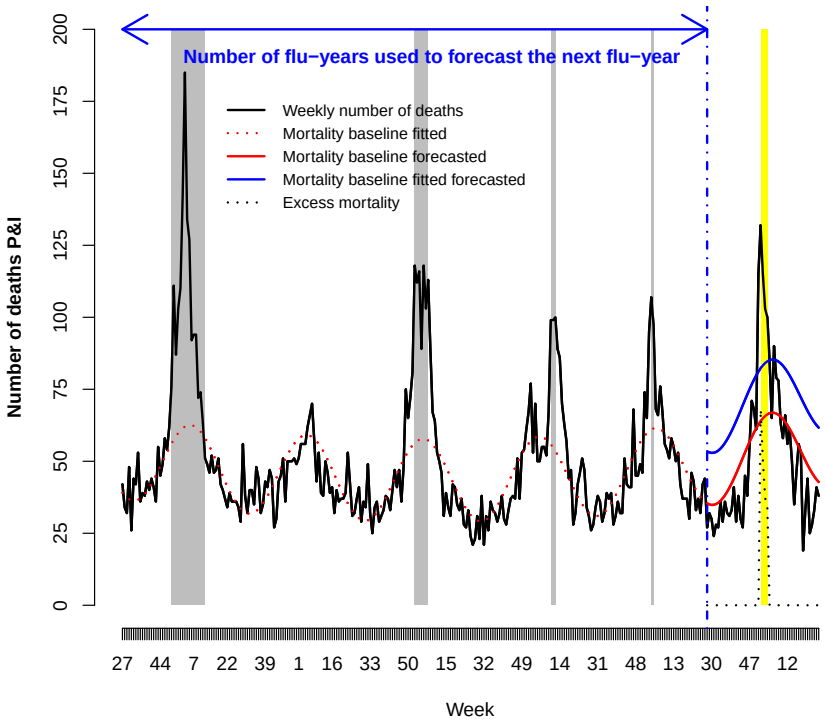

Figure 2.3: Iterative procedure used to fit the statistical model and to identify the $\mathbf{D_a}$ period. Grey boxes represent the E_a periods and yellow boxes represent the D_a.

2.4 General framework

Given the above description it is possible to accommodate a large number of methods in a wide framework of methods varying according to the three points considered, model fitting procedure, type of model and period where to identify the excess mortality periods attributable to the influenza epidemics D_a - see Table 2.1.

Name	Alias	Model	Period	Fitting procedure	T
`It_RM_F`	[15]	Regression	Fixed period	Iterative	5
`It_RM_E`	[14]	Regression	ISS	Iterative	5
`It_SA_F`	none	SARIMA	Fixed period	Iterative	all previous years
`It_SA_E`	[13]	SARIMA	ISS	Iterative	all previous years
`RM_F`	[18, 16]	Regression	Fixed period	Non iterative	NA
`RM_E`	[17]	Regression	ISS	Non iterative	NA
`SA_F`	none	SARIMA	Fixed period	Non iterative	NA
`SA_E`	none	SARIMA	ISS	Non iterative	NA

Table 2.1: Classification of the proposed methods for comparison, according to the fitting procedure (iterative or not), the model (seasonal ARIMA or cyclic regression) and the E_a periods (fixed period or a period estimated by the national Influenza Surveillance Systems, ISS). T represents the dimension of the training set.

Note that it was possible to further identify three new methods, never considered before, from taking all possible combinations of alternatives:

1. `It_SA_F` iterative, using the the seasonal ARIMA model with E_a periods fixed;

2. `SA_F` and `SA_E` that apply the seasonal ARIMA models with a non iterative procedure.

The `It_SA_F` does not seem to present any practical application problems. For methods `SA_F` and `SA_E` we propose to adjust a cyclic regression model to the interrupted time series, substitute then the E_a periods by the model expectations, and then apply the seasonal ARIMA models to this new series.

In order to be able to compare all the methods described in Table 2.1 we had to address the following difficulties:

1. Evaluation of the adjustment quality of the baseline $\beta_{t,a}$ to the original time series $y_{t,a}$ for $(t, a) \notin D_a$. Given that the objective is to estimate a baseline free of excess mortality, the model is not fitted to the observed data during these D_a periods. This condition makes unfeasible the direct application of the more used goodness of fit measures (AIC, BIC, etc.), since the D_a periods and the resulting baseline differ from method to method. To overcome this we have considered the following measures for the baseline outside the D_a:

- Residual Mean Square

$$RMS = \sum_{(t,a)\notin D_a} \frac{\hat{e}_{t,a}^2}{m}$$

where m is the number of weeks without excess deaths attributable to influenza $\hat{e}_{t,a} = y_{t,a} - \beta_{t,a}$ for $(t, a) \notin D_a$;

- Autocorrelation function of the residuals;

2. The true number of excess deaths attributable to influenza epidemics for each flu-year a is unknown. Therefore, the evaluation of the methods considering their excess deaths estimates was accomplished by:

   - comparing, in each flu-year a, the estimated excess deaths attributable to influenza epidemics given by each method, and identifying those that estimate the higher and lower values of excess deaths;
   - calculating the correlation between the estimates of the influenza excess deaths, obtained by different methods, as an empirical concordance measure.

2.5 Application example

The analyzed data consists on the weekly number of deaths by P&I in Portugal from 1980-81 to 2003-04 flu-years obtained from the National Mortality database of the Portuguese Statistics Institute (Figure 2.4).

As presented in Table 2.2 the E_a periods were either set as fixed periods (from week 48 (December) to week 17 (April)) or equal to the non fixed epidemic periods that were defined as follows:

- From 1980-81 to 1989-90 the influenza epidemic periods were defined using the weekly number of deaths by influenza (ICD 9th Revision 487). These periods were set as the consecutive weeks (more than two) with the number of deaths above the 95 percentile of the empirical distribution of the weekly number of deaths by influenza in the period comprised by the flu-years 1980-81 to 1989-90;

- For the flu-years from 1990-91 to 2003-04 the epidemic periods used were the ones defined by the Portuguese ISS, of the Instituto Nacional de Saúde

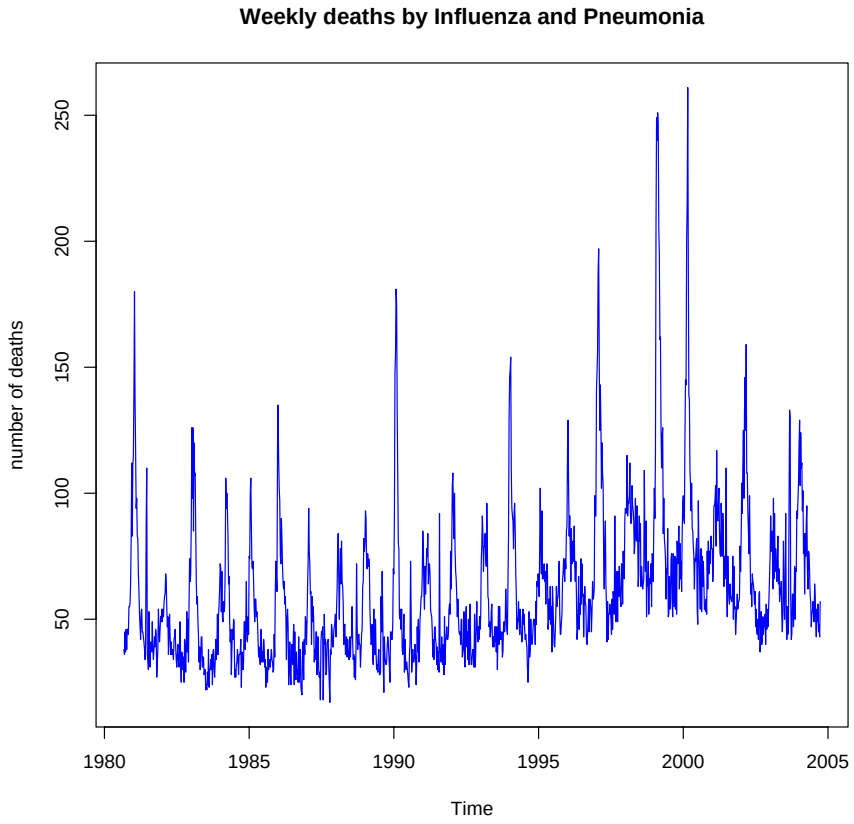

Figure 2.4: Distribution of the weekly number of deaths by influenza and pneumonia in Portugal from 1980-81 to 2003-04.

Dr.Ricardo Jorge (of National Health Institute Dr. Ricardo Jorge) [41]. The final classification is presented in Table 2.2.

In this time series we have also substituted the known heat-waves periods [42] by the average of the number of deaths in the last week before and the first week after the heat-wave.

The models used in the application example were chosen in the following way:

1. The cyclic models used were chosen by a regular best fit model criteria, in a preliminar analysis.

   - non iterative procedure:

$$y_s = \alpha + \beta_1 s + \beta_2 s^2 + \beta_3 s^3 + b_1 \sin\frac{2\pi s}{52} + b_2 \cos\frac{2\pi s}{52} + \varepsilon_s$$

	Epidemic periods (weeks)				
flu-year	**start**	**peak**	**end**	**n° weeks**	**max incidence**
1980-81	49	3	12	16	NA
1981-82	-	-	-	-	NA
1982-83	1	2	7	7	NA
1983-84	10	11	12	3	NA
1984-85	3	3	4	2	NA
1985-86	52	3	3	4	NA
1986-87	-	-	-	-	NA
1987-88	-	-	-	-	NA
1988-89	-	-	-	-	NA
1989-90	1	3	6	6	NA
1990-91	7	9	11	5	148.4
1991-92	45	52	5	13	92.4
1992-93	6	11	14	9	117.7
1993-94	46	49	1	8	168.8
1994-95	3	5	8	6	84.1
1995-96	42	44	51	10	86.8
1996-97	47	50	8	15	111.3
1997-98	-	-	-	-	-
1998-99	51	3	8	10	252.9
1999-00	2	5	8	7	156.5
2000-01	-	-	-	-	-
2001-02	1	5	10	10	239
2002-03	48	50	50	3	76.1
2003-04	44	47	52	9	166.7

Table 2.2: Definition of the epidemic periods E_a for the *flu-years* 1980-81 to 2003-04 (NA: not available). Incidence values are presented by 10^5 inhabitants. From 1980-81 to 1989-90 epidemic periods were defined by the influenza death cause criterium, from 1990-91 to 2003-04 the epidemic periods were defined by the Influenza Surveillance System.

- iterative procedure:

$$y_s = \alpha + \beta_1 s + \beta_2 s^2 + b_1 \sin\frac{2\pi s}{52} + b_2 \cos\frac{2\pi s}{52} + \varepsilon_s$$

2. The seasonal ARIMA models used were:

- non iterative procedure: chosen by an automatic model identification algorithm [43];
- iterative procedure: a model analogous to the one proposed in [13], $ARIMA(2,0,0)(1,1,0)_{52}$:

$$(1 - \phi_1 B - \phi_1 B^2)(1 - \Phi_1 B^{52})\nabla^{52} y_s = \varepsilon_s,$$

where $s = t + (a-1)52$; $t = 1, ..., 52$; $a = 1, ..., 24$; $\varepsilon_s \sim N(0, \sigma^2)$.

In the present application example the D_a periods were always defined as the set of weeks, contained in the E_a periods, that initiate with two weeks with a number of deaths by P&I above the upper 95% confidence limit of the baseline and end with two weeks with a number of deaths bellow the same upper limit.

2.5.1 Results

The methods that considered the E_a periods as fixed presented lower RMS than the methods using year-variable size periods, estimated by the ISS. Within each of these two different groups (fixed period or ISS periods) it was observed that the seasonal ARIMA model always presented lower values of RMS when compared to the cyclic regression model – Figure 2.5. Autocorrelation in the residuals outside of the D_a periods was observed for all the methods that used cyclic regression models.

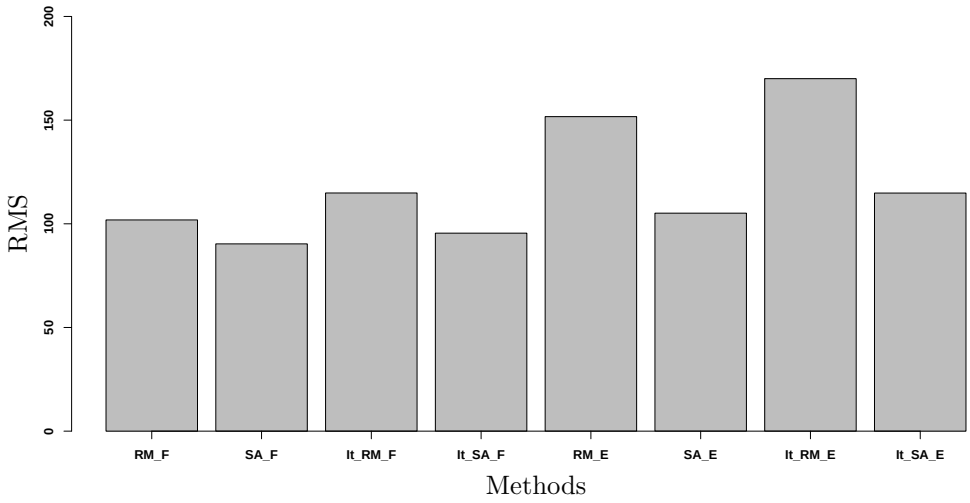

Figure 2.5: Residual Mean Square Errors of the studied models.

From Figures 2.6 and 2.7 it can be observed that when the E_a period is set fixed, the number of influenza-associated estimated deaths is clearly higher. On the other hand, when the E_a periods are set by the ISS it is patent an higher uniformity in the number of influenza-associated estimated deaths, between the methods. This observation was confirmed by the high correlation coefficients obtained between all methods for those using E_a defined by the ISS, that as can be seen in Table 2.3 are all above 0.95. Analyzing the results produced by the iterative approach one can identify

a greater disagreement between estimates after flu year 1993-1994 mainly when the E_a period was fixed.

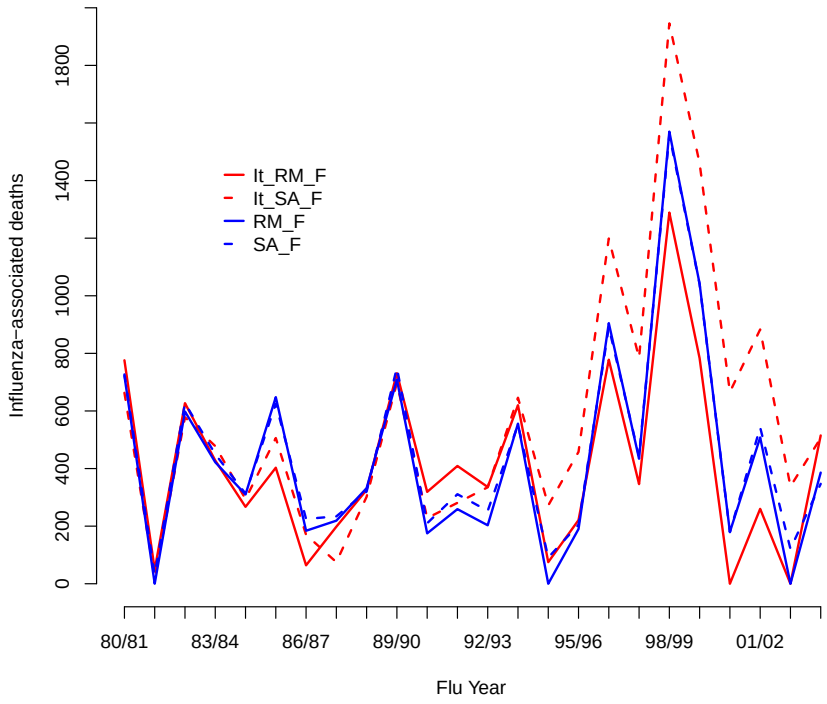

Figure 2.6: Estimated influenza-associated deaths from 1980-81 to 2003-2004 according to the type of method, considering E_a as a fixed period from week 48 (December) to week 17 (April).

	RM_F	SA_F	It_RM_F	It_SA_F	RM_E	SA_E	It_RM_E	It_SA_E
RM_F	1	0.996	0.935	0.909	0.952	0.938	0.914	0.909
SA_F	-	1	0.933	0.914	0.961	0.946	0.917	0.915
It_RM_F	-	-	1	0.793	0.924	0.926	0.932	0.915
It_SA_F	-	-	-	1	0.885	0.872	0.831	0.854
RM_E	-	-	-	-	1	0.996	0.975	0.979
SA_E	-	-	-	-	-	1	0.984	0.986
It_RM_E	-	-	-	-	-	-	1	0.976
I_SA_E	-	-	-	-	-	-	-	1

Table 2.3: Correlation between the estimates of the influenza excess deaths, obtained by different methods.

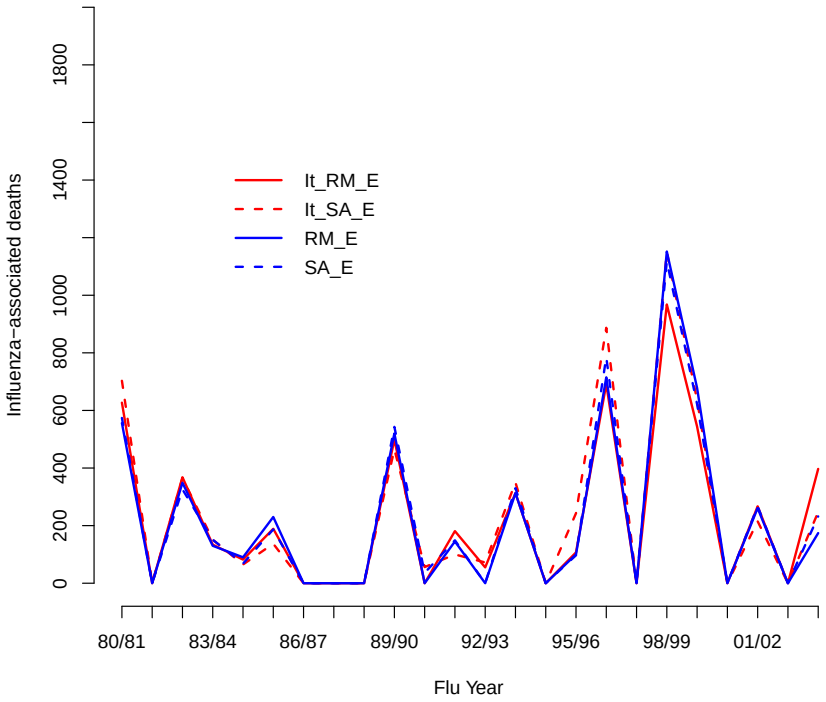

Figure 2.7: Estimated influenza-associated deaths from 1980-81 to 2003-2004 according to the type of method, considering E_a as defined by the Influenza Surveillance System.

2.6 The R-package flubase

The R-package flubase [27] is a set of functions designed to estimate a mortality, or other indicator, baseline free of influenza epidemics, or other time defined event, and the respective attributable excesses for one or more time series (e.g. age groups, gender, regions, etc). The methods available in flubase are the ones described in section 2.4, namely:

1. non iterative with baseline estimated by a cyclical regression (RM_F and RM_E) implemented in the function: baseRM;

2. non iterative with baseline estimated by ARIMA models (SA_F and SA_E) implemented in the function: baseSA;

3. iterative with baseline forecasted by a cyclical regression (It_RM_F and It_RM_E) implemented in the function: baseIt_RM;

4. iterative with baseline forecasted by ARIMA models (It_SA_F and It_SA_E) implemented in the function: baseIt_SA.

Additionally to the specific methods functions it was also implemented a general function:

flubase(dat, groups, per, pe = 0, method, indicator=mortality, g_label)

where,

- dat: a data.frame with all the variables needed: dat$group indicating the group to each time series belongs, dat$year indicating the civil year, dat$todeath is the time unit index that can be week or month, dat$nod are the number of deaths observed at each time and dat$epi is an indicator variable of the epidemic period (dat$epi=1 if the week or month belongs to the epidemic period and dat$epi=0 otherwise);

- groups: number of groups considered, e.g. number of age groups, regions, etc.;

- per: parameter that defines if data is weekly (per = 52) or monthly (per = 12);

- pe: parameter that states if the user will provide the epidemic periods or uses fixed periods. pe = 0 if the user provides the epidemic periods in the epi parameter; otherwise, pe = 1 if the function uses a fixed period from week 47 to week 17 or from month 12 to month 4;

- method: the method used to estimate the baseline, method=c("nrm","nsa","irm","isa"). nrm: non iterative multiple regression, nsa: non iterative seasonal ARIMA, irm: iterative multiple regression and isa: iterative seasonal ARIMA;

- `indicator`: a text string indicating the name of the indicator. By default indicator="mortality";
- `g_label`: a vector string whose length is the number of groups, containing the labels for the groups, e.g. `g_label` = `c`($male$","$female$")" or `g_label` = `c`($0-14yrs$","$15-44yrs$","$45-64yrs$","$65+yrs$")".

This function can apply any of the framework methods to a pre-prepared data set `dat` producing a text file with the vectors, `beta0`, the baseline vectors, `beta_up`, the upper 95% confidence limit of the baseline and, `da`, a vector of 0 and 1s indicating the weeks or months with excess deaths attributable to an influenza epidemic. Additionally `flubase` produces a graph showing the observed number of deaths, the baseline, the respective upper 95% confidence interval, the epidemic periods and the periods with excess deaths attributable to influenza epidemics, in grey and yellow boxes respectively.

2.6.1 Example

For illustrative purposes consider the time series of weekly number of deaths by all causes for Portugal in the period from 1997 to 2004.

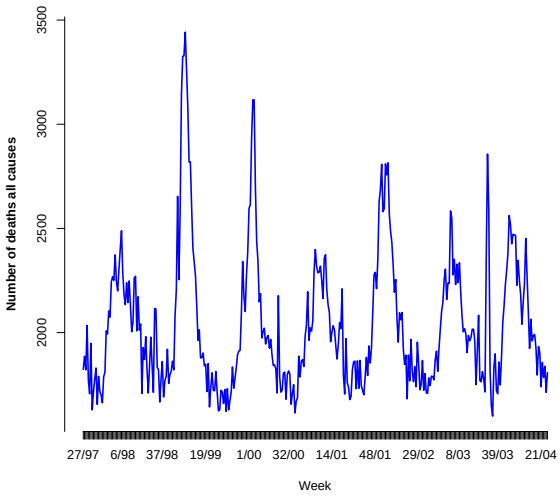

Figure 2.8: Weekly number of deaths from all causes in Portugal for the period from 1997 to 2004

So, in order to estimate excess deaths associated with influenza epidemics using the `flubase` package, the following situations are presented:

- E_a periods as fixed and the non iterative procedure with a cyclical regression model (Figure 2.9 and Table 2.4) `>flubase(data, groups=1, per=52, pe = 1, method="nrm", indicator = "mortality", g_label="all causes")`.

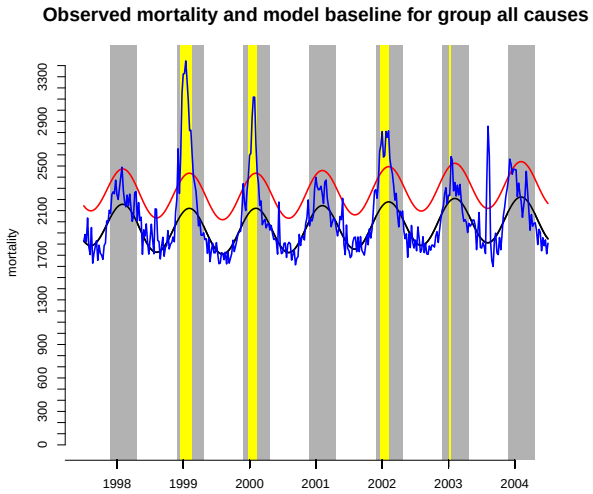

Figure 2.9: Output of `flubase` package considering E_a periods as fixed and the non iterative procedure with a cyclical regression model. Blue line is the observed number of deaths, black line is the baseline, red line is the upper 95% confidence limit for the baseline, grey boxes are the E_a periods, yellow boxes are the D_a periods.

Event	Observed	Expected	Excess
1	-	-	-
2	30,410	21,016	9,394
3	16,910	12,681	4,229
4	-	-	-
5	18,933	15,165	3,768
6	5,135	4,399	736
7	-	-	-

Table 2.4: Output of `flubase`: excess deaths estimates for each E_a period, considering E_a periods as fixed and the non iterative procedure with cyclical regression model.

- E_a periods provided by the user and the non iterative procedure with a seasonal ARIMA model (Figure 2.10 and Table 2.5) `>flubase(data, groups=1, per=52, pe = 0, method="nsa", indicator = "mortality", g_label="all causes")`

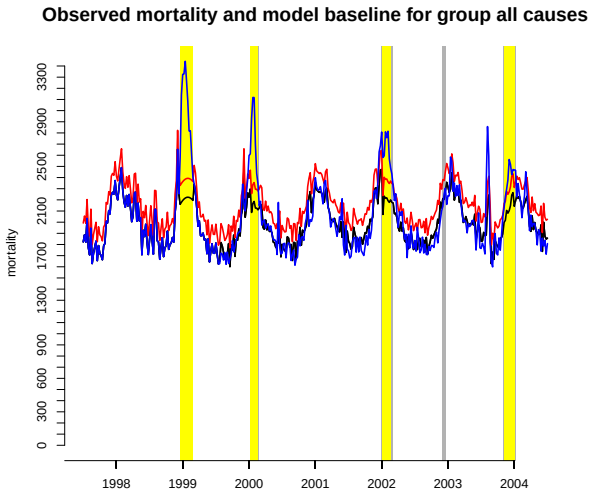

Figure 2.10: Output of `flubase` package considering E_a periods provided by the user and the non iterative procedure with a seasonal ARIMA model. Blue line is the observed number of deaths, black line is the baseline, red line is the upper 95% confidence limit for the baseline, grey boxes are the E_a periods, yellow boxes are the D_a periods

Event	Observed	Expected	Excess
1	32,815	24,244	8,571
2	19,254	15,100	4,154
3	23,851	19,956	38,95
4	-	-	-
5	21,820	19,298	2,522

Table 2.5: Output of the `flubase`: excess deaths estimates for each E_a period, E_a periods provided by the user and the non iterative procedure with a seasonal ARIMA model.

As presented before the methods implemented in `flubase` were developed for the purpose of estimating excess deaths attributable to influenza epidemics. Nonetheless given the generality of the propose, the mortality indicator can be replaced by other indicators, like hospitalizations, medical consultations, drug sales, etc. On the other hand, we can go even further on this generalization by considering other events with impact on mortality or other indicators.

As can be seen in Figure 2.8 there is a high peak of mortality during the summer of 2003. This excess of mortality was already shown to be associated with a heat-wave occurred from 30 July to 15 August 2003 [44]. So if one wanted to use `flubase` package to estimate this observed excess one would need to set the weeks where the heat-wave occurred as an E_a period. In this approach the E_a periods are generalized to a period where an event (epidemic, environmental, etc) with potential impact on the indicator of interest (deaths, hospitalizations, consultations, etc) has occurred. In this context, after setting the weeks of the heat-wave as 1 in the `data` file used in the previous example one can the call `flubase` function as follows (see results in Figure 2.11 and Table 2.6):

```
>flubase(data, groups=1, per=52, pe = 0, method="nrm", indicator = "mortality",
g_label="all causes")
```

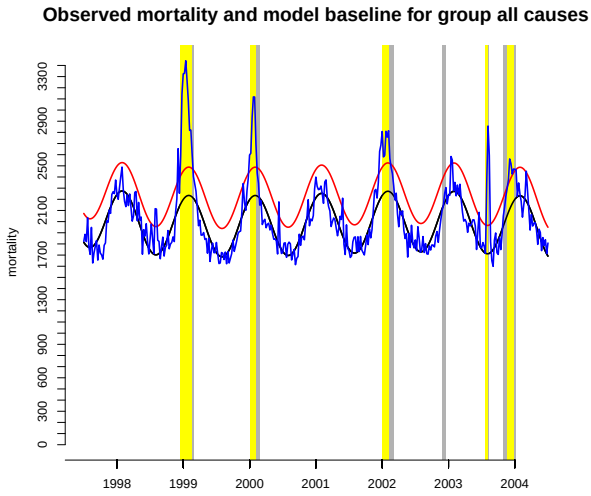

Figure 2.11: Output of `flubase` package considering E_a periods provided by the user (including the 2003 heat-wave) and the non iterative procedure with a cyclical regression model. Blue line is the observed number of deaths, black line is the baseline, red line is the upper 95% confidence limit for the baseline, grey boxes are the E_a periods, yellow boxes are the D_a periods

As can be seen from Table 2.6 the `flubase` estimated an excess of 2,665 deaths attributable to the 2003 heat-wave, using the non interactive cyclical regression model. This estimate do not greatly differ from the previously published estimates [44] that vary from 1,884 to 2,229, obtained with other methods and with daily data.

Event	Observed	Expected	Excess
1	30,410	22,126	8,284
2	14,473	11,131	3,342
3	18,933	15,822	3,111
4	-	-	-
5	7,804	5,139	2,665
6	14,827	12,877	1,950

Table 2.6: Excess deaths estimates for each E_a period (including the 2003 heat-wave), E_a periods provided by the user and the non iterative procedure with a cyclical regression model.

2.7 Discussion

The present work constitutes an important step towards a unified analysis of the methods that use interrupted mortality time series to estimate death associated with influenza epidemics. This was achievement via the identification of the principal similarities and differences between the main existing methods, leading to the definition of a methodological class and a subsequent parametrization of its members in terms of the type of statistical model, the a priori chosen type of period to estimate the epidemic period and the procedure employed to fit the model to the interrupted mortality time series and identify the periods with excess deaths associated to influenza epidemics. It is also important to notice that using this very broad structure three new methods, never considered before, were identified and proposed: iterative approach with seasonal ARIMA model considering E_a periods as fixed and the non-iterative approach using the seasonal ARIMA model considering E_a periods as fixed or as estimated from an external data source.

Additionally, another important output of this parametrization is a set of `R`-routines, package `flubase`. These user friendly routines, downloadable at `http://cran.r-project.org/web/packages/flubase/index.html` [27], constitute an important tool for the estimation of the influenza associated deaths, as one can easily compare the results obtained by varying each one of the parameters involved. Further, this package can also be used to estimate excesses of deaths or of other indicators, attributable to other events defined in time and with potential impact on the indicator of interest. For instance the non iterative cyclical regression model was already applied with success to estimate excess deaths attributable to heat-waves [45, 46]. In this case, the E_a period was set as the number of weeks where the heat-wave occurred.

Pelat et al [47] have also developed a tool that is able to automatically obtain a baseline and quantify the excess of deaths attributable to influenza allowing the user to vary some parameters. Nevertheless, they have only included cyclical regression models and have not made comparisons between the results obtained with the different methodological approaches.

We must mention that not all the differences between methods were parameterized. One that was not considered was the possible replacement of the values in the interrupted periods of the time series by the corresponding forecasts obtained from a training set period. Note that here, we have not replaced the values in the interrupted periods when the model chosen was the cyclic regression one but, when the model was the ARIMA, we have substitute them by previous years forecasts, in accordance with the original paper where they have been proposed. This option was taken for simplicity sake.

We note that, in the following discussion, the comparisons between methods are only supported by a single application, the weekly number of deaths by P&I in Portugal. In our opinion, this fact does not pose a problem as the inconsistencies observed in the results between methods are enough to question the idea that they all produce equivalent results.

As mentioned before, for comparing methods the main question is not which one will produce the best results. Since the true number of influenza associated deaths is unknown we have to rely on relative comparisons of results between methods and associated class parameters.

In this sense, we have observed that the most sensitive parameter was the type of a priori chosen E_a period. As should be expected, the higher estimates were obtained when this period was set fixed from December to April. Intuitively, one can say that this type of period is more sensitive and that the one defined by the ISS gives a more specific estimate but also more conservative. In our opinion, this parameter is the cornerstone in establishing the epidemiological association between the occurrence of influenza epidemics and the simultaneous observation of excess of deaths in the mortality time series. We have further obtained a lower RMS when the E_a period was fixed, which can be explained by the fact that in this situation a great part of the remaining time series belongs to spring and summer where the observations variation is lower.

Considering the type of model fitted we have seen that the seasonal ARIMA

model presents always lower RMS than the cyclic regression model and also non autocorrelated residuals off the D_a periods. The main consequence of these results is a lower upper 95% confidence limit for the methods using seasonal ARIMA models. This gain in efficiency has been also reported in other works [13].

The baseline building process also seems to be a parameter with important impact on the estimates. In this application we have seen that the estimates from the iterative procedure differ from the estimates of the non-iterative procedure essentially after the 1993-94 flu-year. From Figure 2.4 we can see that this is the moment when the time series presents an important change in trend. Generally speaking, the iterative processes are more likely to produce estimates further apart from the observed values than the estimates obtained by fitting the model directly to the entire time series, because these methods only use data from previous flu-years to estimate each flu-year baseline. This is more evident where, as it happens in here, the iterations are made with weekly observations for an entire year ahead. In fact, the first proposals of the iterative approach in the literature were not to be applied to this problem but to build baselines for surveillance purposes [13, 48], that makes all the sense since the next flu-year data was not yet observed. To be true, this is the method used in the actual mortality surveillance systems [8, 9, 45].

Chapter 3

Excess mortality associated with influenza epidemics in Portugal, 1980 to 2004[1]

3.1 Introduction

As referred previously, seasonal influenza epidemics have a substantial mortality and morbidity impact on human health globally [49, 50, 51]. In the US the most recent estimates, provided by the CDC, for the period from 1976 to 2007, point to a seasonal average of 23,607 deaths ranging from 3,349 in 1986-87 to 48,614 in 2003-04 [55]. For Europe the European Center for Prevention and Disease Control (ECDC) considers that the overall seasonal influenza epidemics burden is about 40,000 deaths [6].

The exact burden of influenza is difficult to quantify because laboratory tests are rarely conducted on a routine basis. Further, influenza can trigger secondary bacterial infections or exacerbate existing chronic conditions, which can lead to hospitalization or death, even after the primary viral infection has been cleared. As a result, as it was discussed in Chapter 2, influenza disease burden studies rely on the application of statistical time series methods to broadly-defined disease outcomes, such as mortality and hospitalization from P&I or respiratory and cardiovascular diseases, or all-cause mortality to estimate the excesses attributable to the influenza epidemics [16, 33, 52, 53, 54].

[1]This Chapter is based on the paper Nunes B, Viboud C, Machado A, Ringholz C, Rebelo-de-Andrade H, et al. (2011) Excess Mortality Associated with Influenza Epidemics in Portugal, 1980 to 2004. PLoS ONE 6(6): e20661. doi:10.1371/journal.pone.0020661

A substantial body of evidence suggests that age is one of the most important risk factors when considering the health impact of influenza. Children under 5 years and adults of 65 years and older are considered to be at an increased level of risk for influenza-related complications during inter-pandemic periods [33, 56]. Several studies have explored the influenza-associated rates of hospitalization and death among seniors in the US, Canada, and Europe, as well as in a few high-income tropical settings [18, 34, 57, 58, 59].

Very little information, however, is available for Southern Europe, with only one mortality study set in Italy [18]. In particular, no estimates for Portugal exist in the English-speaking literature, and the ones available corresponds to all causes mortality outcome for all ages and for the elderly ($\geq$ 65 years) in the period from 1990 to 1998. In this study the authors estimated a seasonal average of 1,774 deaths attributable to influenza epidemics raging from 806 to 3,979 [21].

Given this lack of knowledge for Portugal and for the South of Europe regarding the influenza epidemic burden, its this study aim to provide estimates of excess mortality associated with influenza virus activity in Portugal by death category (cerebrovascular disease, ischemic heart disease, diseases of the respiratory system, chronic respiratory diseases, and pneumonia and influenza), age group (0-4, 5-54, 65-69, 70-74, 75-79, 80-84 and $\geq$ 85), and circulating virus subtypes for 1980-2004 and compare our estimates with those from other locations.

3.2 Data

3.2.1 Mortality and population data

Mortality data available for 1980-2004 were obtained from the Portuguese National Mortality Database of the Instituto Nacional de Estatística (National Statistics Institute in Portugal). The 9th revision of the WHO International Classification of Diseases (ICD-9) was used until 2001, and the 10th revision (ICD-10) was used thereafter. Mortality time series data were compiled monthly according to age group (0-4, 5-54, 55-64, 65-69, 70-74, 75-79, 80-84 and $\geq$ 85 years) for all-cause mortality and the following primary causes of death: cerebrovascular disease (ICD-9: 430-438, ICD-10: I60.0-I69.8); ischemic heart disease (ICD-9: 410-414, ICD-10: I20.0-I25.9); diseases of the respiratory system (ICD-9: 460-519, ICD-10: J00-J99); chronic respiratory diseases, including bronchitis, asthma, and emphysema (ICD-9: 490-495 and ICD-10:

J41-J47); pneumonia (ICD-9: 480-486, ICD-10: J12.0-J18.9) and influenza (ICD-9: 487, ICD-10: J10.0-J11.8).

To evaluate the specificity of our approach, we also studied deaths from intentional and non-intentional injuries (ICD-9, E-codes and ICD-10, codes V-Y), which we considered as a "control" outcome with no direct causal association with influenza activity.

In order to cover influenza epidemic seasons, which can occur from October to May in Portugal, 24 respiratory seasons were defined starting in July of the first available year (1980) and ending in June of the last available year (2004).
So, for each of these causes of death, let us consider $y_{t,a,g}$ the number of observed deaths for month $t = 1, ..., 12$, flu-year $a = 1(1980/81), ..., 24(2003/04)$ and age $g = 1(0-4), ..., 8(\geq 65)$. Age-specific annual population estimates from 1980 to 2004 were downloaded from the Instituto Nacional de Estatística website (http://www.ine.pt) [60] and used to derive monthly death rates by disease outcome and age group. Additionally, all monthly time series were standardized to a fixed number of days in the month (30.4 days) and the time series of the mortality rates per inhabitant were calculated:

$$r_{t,a,g} = \frac{y_{t,a,g}}{N_{a,g}} \frac{30.4}{m_t}$$

were m_t represents the number of days of month t and $N_{a,g}$ the total population estimate in flu-year a and age group g.

3.2.2 Influenza-like illness and virological surveillance data

An integrated clinical, epidemiological and virological influenza surveillance system was established in Portugal in the winter of 1990-1991. Data on ILI, as defined in primary care [61], were collected by a network of Sentinel GP's (Médicos-Sentinela) which total patient list represents approximately 2.3% of the Portuguese population. Weekly ILI incidence rates were calculated based on these data. A subset of 25-35% of GPs also provided respiratory specimens to conduct virological surveillance. Respiratory specimens were centralized at the Centro Nacional da Gripe (National Influenza Centre) of Instituto Nacional de Saúde Dr. Ricardo Jorge in Lisbon and tested for influenza A/H3N2, A/H1N1 and B subtypes by PCR or culture. Influenza subtype dominance each season was defined as the subtype that was isolated in at least 51% of the influenza-positive ILI cases.

3.3 Definition of epidemic periods E_a

For most of the study period, 1980-81 to 2001-02, influenza epidemic periods E_a were estimated based on monthly deaths for all ages coded specifically as influenza, which is considered a specific indicator of the timing of epidemics [16, 18]. We used a Serfling [48] like approach i.e. a non interactive cyclical regression model with E_a periods fixed (December to April) to predict baseline mortality in the absence of influenza epidemics and define epidemic periods, as in [16, 18]. The model included time trends and seasonal terms as following:

$$y_s = \alpha + a_1 s + a_2 s^2 + a_3 s^3 + b_1 \sin\frac{2\pi s}{52} + b_2 \cos\frac{2\pi s}{52} + \varepsilon_s$$

where s is a running index for time $s = t + (a-1)12$ for month $t = 1, ..., 12$ and flu-year $a = 1, ..., 24$; y_s is the number of influenza-specific deaths in month s; α is the intercept; a_i are coefficients for time trends in baseline mortality; b_1 and b_2 are seasonality coefficients; and ε_s represents normally distributed errors.

The estimates of the epidemic periods E_a were then defined as the set of consecutive months from November to April where the observed number of influenza-specific deaths exceeds the 95% upper confidence limit of the model baseline [16, 18].

Following the transition from the 9th to the 10th revision of ICD in 2002, influenza-specific mortality decreased dramatically, and influenza epidemic periods could not be defined based on these data alone. For the remaining seasons 2002-2004, epidemic periods were defined by the Portuguese ISS as the weeks where ILI incidence rate exceeded the upper 95% confidence limit of the ILI baseline and where influenza virus circulation was confirmed by laboratory tests, as in [62]. Since mortality data were available on a monthly basis, we considered a given month to be epidemic if at least one week during that month was considered epidemic in the ISS. We also checked the consistency of using influenza-specific deaths and ILI incidence to estimate the E_a periods, based on the subset of years when both types of data were available (1991-2002).

3.4 Estimation of influenza-associated excess deaths

In order to estimate the excess deaths associated with influenza epidemics in Portugal for each age group and death outcome (pneumonia, chronic respiratory diseases,

all respiratory diseases, cerebrovascular diseases, ischemic heart diseases) one of the methods from the general framework described in section 2.4 was chosen. Hence, in the context of the results obtained in Chapter 2 the seasonal ARIMA model with a non iterative fitting procedure (for details please see Section 2.4) and E_a periods estimated as described in previous section was the selected method. This decision was based on the fact that 1. Seasonal ARIMA models are more efficient in handling time series with auto-correlation, 2. The non iterative procedure is the most adequate to retrospectively estimate influenza-associated excess deaths and 3. using E_a periods estimated from external sources of data is a more conservative approach which is more appropriate given the ecologic nature of this study.

The chosen method was then applied to the natural logarithm of death rates $z_{t,a,g} = \ln(r_{t,a,g})$. This data transformation was performed in order to stabilize the time series variance.

In order to evaluate model adequacy, we computed the Box-Ljung test for auto-correlation and the Kolmogorov-Smirnov test for normality of model residuals outside of epidemic periods.

The obtained baseline and the respective upper 95% confidence limit of the log-death rates were then exponentiated to estimate the $D_{a,g}$ periods, i.e. the periods with excess death rates attributable to influenza epidemics for each flu-year a and age group g. More specifically the baseline of the death rates in the absence of influenza epidemics and the respective upper 95% confidence limit were calculated by:

$$\hat{\beta}^r_{t,a,g} = \exp\left(\hat{\beta}^z_{t,a,g}\right)$$

$$u(\hat{\beta}^r_{t,a,g}) = \exp\left(\hat{\beta}^z_{t,a,g} + p_{0.975}\hat{\sigma}_{e,g}\right)$$

were $\hat{\beta}^z_{t,a,g}$ is the estimated baseline of the natural logarithm of the death rates without the effect of the E_a periods predicted by the ARIMA model, $\hat{\sigma}_{e,g}$ is the estimated standard deviation of the seasonal ARIMA model and $p_{0.975}$ is the 97.5% percentile of the standard normal distribution.

The $D_{a,g}$ periods were then estimated as the consecutive months contained in the E_a periods where $r_{t,a,g} > u(\hat{\beta}^r_{t,a,g})$.

For the months included in the $D_{a,g}$ periods the excess death rates attributable

to influenza epidemics were then estimated as:

$$r^{ex}_{t,a,g} = r_{t,a,g} - \hat{\beta}^{r}_{t,a,g}$$

and the absolute number of excess deaths as:

$$y^{ex}_{t,a,g} = r^{ex}_{t,a,g} A_{t,a,g},$$

where $A_{t,a,g} = \frac{m_t N_{a,g}}{30.4}$.

Given that our main aim was to estimate the total excess deaths attributable to each occurred influenza epidemic, corresponding to each E_a period, one needs to obtain first the total excess deaths in each E_a period for each age group:

$$y^{ex}_{a,g} = \sum_{(t,a)\in D_{a,g}} y^{ex}_{t,a,g} \tag{3.1}$$

and finally the total excess deaths in the E_a for all ages groups:

$$y^{ex}_{a} = \sum_{g=1}^{8} y^{ex}_{a,g} \tag{3.2}$$

The excesses attributable to influenza epidemics are presented as absolute numbers, and as crude and age-standardized rates using the 2000 world population as a reference [63]. They are calculated as:

$$rst^{ex}_{a} = \sum_{g=1}^{8} w_g \frac{y^{ex}_{a,g}}{N_{a,g}} \tag{3.3}$$

where w_g is the weight of age group g in the world population estimated for year 2000. Additionally the proportion of deaths attributable to influenza among all deaths occurring from October to May (influenza season) is also presented, averaged across all influenza seasons.

3.4.1 Excess deaths confidence intervals

Confidence intervals for seasonal excess death and age-standardized excess death rates were obtained by taking into account the uncertainty of the baseline and assuming that the sum of monthly excess deaths during epidemic periods followed a log-normal distribution.

In order to obtain a confidence interval for the absolute number of deaths attributable to influenza epidemic E_a (y_a^{ex}) we start by finding the distribution $y_{a,g}^{ex}$.

From equation 3.2 one can assume that:

$$y_{a,g}^{ex} = \sum_{(t,a)\in D_{a,g}} (y_{t,a,g} - \hat{\beta}_{t,a,g}^{y})$$

where

$$\hat{\beta}_{t,a,g}^{y} = \exp(\hat{\beta}_{t,a,g}^{z}) A_{t,a,g},$$

is the expected number of deaths, in the absence of influenza epidemics, for month t flu-year a and age group g. Admitting that the observed deaths ($y_{t,a,g}$) are fixed, we only need to find the distribution of expected number of deaths according to the baseline $\hat{\beta}_{t,a,g}^{y}$.

From the ARIMA model one can consider that $\hat{\beta}_{t,a,g}^{z} \sim N(x_{t,a,g}, \sigma_{e,g}^2)$, conditional on $x_{t,a,g}$, where $x_{t,a,g}$, like in equation 2.1, is the time series of the log-transformed rates with E_a periods imputed by the predictions ($\tilde{z}_{t,a,g}$) obtained from the cyclical regression models fitted to the interrupted time series $z_{t,a,g}^{\star} = \{z_{t,a,g} : (t,a,g) \notin E_a\}$:

$$x_{t,a,g} = \begin{cases} z_{t,a,g}, & (t,a) \notin E_a \\ \tilde{z}_{t,a,g}, & (t,a) \in E_a \end{cases}$$

In this context one can assume that:

$$\hat{\beta}_{t,a,g}^{y} = A_{t,a,g} \exp\left(\hat{\beta}_{t,a,g}^{z}\right) \sim \text{Log-N}(\ln A_{t,a,g} + x_{t,a,g}, \sigma_{e,g}^2).$$

Given this result from the Fenton-Wilkinson approximation [64] the sum of the baseline values for each $D_{a,g}$ period will also be Log-Normal:

$$\hat{\beta}_{a,g}^{y} = \sum_{(t,a)\in D_{a,g}} \hat{\beta}_{a,g,t}^{y} \sim \text{Log-N}(\mu_{\hat{\beta}_{a,g}^{y}}, \sigma_{\hat{\beta}_{a,g}^{y}}^2)$$

where

$$\sigma_{\hat{\beta}_{a,g}^{y}}^2 = \ln\left((\exp(\sigma_{e,g}^2) - 1) \frac{\sum_{(t,a)\in D_{a,g}} \exp\left(2\left(\ln(A_{t,a,g}) + x_{t,a,g}\right)\right)}{\left(\sum_{(t,a)\in D_{a,g}} \ln(A_{t,a,g}) + x_{t,a,g}\right)^2} + 1 \right)$$

and

$$\mu_{\hat{\beta}_{a,g}^{y}} = \ln\left(\sum_{(t,a)\in D_{a,g}} \ln(A_{t,a,g}) + x_{t,a,g} \right) + \frac{\sigma_{e,g}^2}{2} - \frac{\sigma_{\hat{\beta}_{a,g}^{y}}^2}{2}$$

Finally, using also the Fenton-Wilkinson approximation, the distribution of the expected number of deaths in the absence of the influenza epidemics effect, for all age groups, is given by:

$$\hat{\beta}_a^y = \sum_{g=1}^{8} \hat{\beta}_{a,g}^y \sim \text{Log-N}(\mu_{\hat{\beta}_a^y}, \sigma^2_{\hat{\beta}_a^y})$$

where

$$\sigma^2_{\hat{\beta}_a^y} = \ln\left(\frac{\sum_{g=1}^{8} \exp(2\mu_{\hat{\beta}_{a,g}^y} + \sigma^2_{\hat{\beta}_{a,g}^y})(\exp(\sigma^2_{\hat{\beta}_{a,g}^y}) - 1)}{\left(\sum_{g=1}^{8} \exp\left(2\mu_{\hat{\beta}_{a,g}^y} + \frac{\sigma^2_{\hat{\beta}_{a,g}^y}}{2}\right)\right)^2} + 1\right)$$

and

$$\mu_{\beta_a^y} = \ln\left(\sum_{g=1}^{8} \exp\left(2\mu_{\hat{\beta}_{a,g}^y} + \frac{\sigma^2_{\hat{\beta}_{a,g}^y}}{2}\right)\right) - \frac{\sigma^2_{\hat{\beta}_a^y}}{2}$$

In summary the 95% confidence interval for the all ages excess deaths attributable to the influenza epidemic E_a (y_a^{ex}) was obtained by subtracting the upper 0.975 and lower 0.025 probability quantiles of the β_a^y distribution from the sum of observed number of deaths for all age groups $y_a = \sum_{g=1}^{8} \sum_{(t,a)\in D_{a,g}} y_{t,a,g}$ in the $D_{a,g}$ periods.

Another measure of interest calculated in this work was the age-standardized excess death rates for each E_a period. Considering that:

$$\begin{aligned} rst_a^{ex} &= \sum_{g=1}^{8} \frac{w_g}{N_{a,g}} y_{a,g}^{ex} = \\ &= \sum_{g=1}^{8} W_g (y_{a,g} - \hat{\beta}_{a,g}^y) \end{aligned}$$

and assuming that $\sum_{g=1}^{8} W_g y_{a,g}$ is known we need to obtain the distribution of $\hat{\beta}_a^{rst} = \sum_{g=1}^{8} W_g \hat{\beta}_{a,g}^y$, where $W_g = \frac{w_g}{N_{a,g}}$.

So according to the previous results:

$$W_g \beta_{a,g}^y \sim \text{Log-N}(\ln(W_g) + \mu_{\hat{\beta}_{a,g}^y}, \sigma^2_{\beta_{a,g}^y})$$

and, from the Fenton-Wilkinson approximation the sum over all age groups:

$$\hat{\beta}_a^{rst} \sim \text{Log-N}(\mu_{\hat{\beta}_a^{rst}}, \sigma^2_{\hat{\beta}_a^{rst}})$$

where

$$\sigma^2_{\hat{\beta}^{rst}_a} = \ln\left(\frac{\sum_{g=1}^{8}\exp\left(2\ln(W_g)+\mu_{\hat{\beta}^{y}_{a,g}}+\sigma^2_{\hat{\beta}^{y}_{a,g}}\right)\left(\exp(\sigma^2_{\hat{\beta}^{y}_{a,g}})-1\right)}{\left(\sum_{g=1}^{8}\exp\left(2\ln(W_g)+\mu_{\hat{\beta}^{y}_{a,g}}+\frac{\sigma^2_{\hat{\beta}^{y}_{a,g}}}{2}\right)\right)^2}+1\right)$$

and

$$\mu_{\hat{\beta}^{rst}_a} = \ln\left(\sum_{g=1}^{8}\exp\left(2\ln(W_g)+\mu_{\hat{\beta}^{y}_{a,g}}+\frac{\sigma^2_{\hat{\beta}^{y}_{a,g}}}{2}\right)\right)-\frac{\sigma^2_{\hat{\beta}^{rst}_a}}{2}$$

To finalize the 95% confidence interval for rst^{ex}_a is obtained by subtracting the upper 0.975 and lower 0.025 probability quantiles of $\hat{\beta}^{rst}_a$ distribution from $\sum_{g=1}^{8} W_g y_{a,g}$. All results were computed in the R Environment for Statistical Computing [28].

3.5 Results

3.5.1 Overall burden of influenza

Portugal experiences highly seasonal influenza activity concentrated in winter months, with peaks in influenza-specific mortality occurring between December and March. As in other developed countries, the impact of influenza epidemics varies greatly between years, as illustrated by important year-to-year variation in the size of P&I and all-cause mortality peaks (Figure 3.1).

During the study period, 1980-2004, the seasonal average number of all cause excess deaths associated with influenza epidemics was 2,475 in Portugal (range=0 to 8,514), 90% of which occurred in people aged $\geq$ 65 years, representing a crude excess all-cause death rate of 24.7 per 100,000. The corresponding average age-standardized rate was 13 per 100,000 inhabitants, representing an average of 3% of total deaths occurring between October and May (range 0 to 9.2%, Table 3.1). In seniors aged 65 years and over, the average age-standardized excess death rate during these months was 156 per 100,000 inhabitants, representing 4% of all October-May deaths in this age group (range 0 to 11.9%). For cerebrovascular and ischemic heart disease outcomes, the average age-standardized death rate were 2.9 and 0.7 per 100,000, respectively, representing averages of 3.2% and 2% of the age-standardized mortality rate for those causes during the influenza season (Table 3.2). The age-standardized mortality rates was 3.1/100,000 for respiratory diseases, 1.5/100,000 for

P&I, and 0.8/100,000 for chronic respiratory diseases. Influenza was responsible for 9.9%, 9.3% and 6.9% on average of all deaths from P&I causes, respiratory diseases, and chronic respiratory diseases, respectively. Similar results were observed for the elderly ($\geq$ 65 years) (Table 3.2). We estimate than on average 662 respiratory deaths are attributable to influenza every season in Portugal, of which 44% are coded as P&I.

3.5.2 Age-specific estimates

The distribution of age-specific all-cause excess mortality rates followed a J-shape, with highest rates in seniors over 65 years, and the 0-4 age group experiencing higher excess mortality rates than individuals aged 5-54 years (Table 3.2 and Figure 3.2). In children under 5 years, no excess deaths could be attributed to influenza in chronic respiratory diseases, ischemic heart diseases or cerebrovascular diseases. There were very few deaths coded as such in this age group. Excess mortality rates increased exponentially with age for age groups over 55 years. This pattern was also observed for excess mortality from P&I and chronic respiratory diseases but not for the other causes of death studied. Interestingly, the J-shape almost disappeared when exploring the age-specific proportion of October-May deaths attributable to influenza, which instead demonstrated a near linear association with age (Figure 3.2). This suggests that measuring the proportion of excess deaths due to influenza is a good way to standardize burden estimates across age groups (and possibly across time and geography).

3.5.3 Burden of influenza according to season and dominant subtype

Of the 24 influenza seasons studied, 14 were dominated by the more severe A/H3N2 subtype (Table 3.1). Average excess mortality rates were 3.3-6.1 higher for seasons dominated by A(H3) viruses compared to seasons dominated by influenza B or A(H1), depending on the outcome studied (Table 3.2). Seasons associated with the highest excess mortality rates (e.g, 1998-99 and 1980-81) had an especially high disease burden in people aged $\geq$ 65 years. Five of the 24 seasons were associated with no excess death in any of the studied causes.

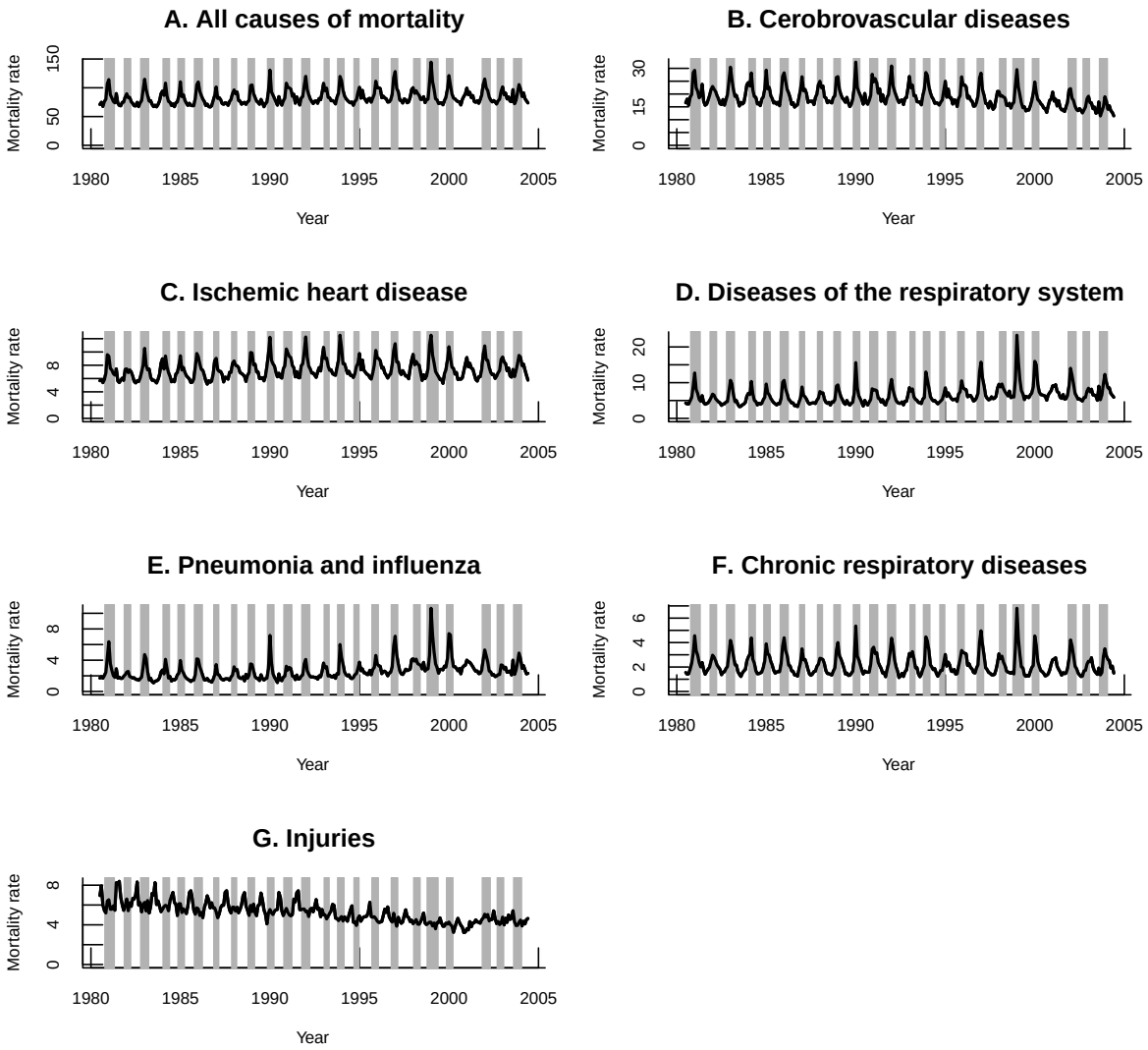

Figure 3.1: All age mortality rates for A. All causes, B. Cerebrovascular diseases, C. Ischemic heart diseases, D. Diseases of the respiratory system, E. Pneumonia and Influenza, F. Chronic respiratory diseases and G. injuries from 1980/81 to 2003/2004 in Portugal. Grey highlights represent influenza epidemic periods.

Season	Epidemic Periods based on ISS	Epidemic Periods based on ISM	Number of epidemic months ISM	Dominant virus (sub)type**	All causes influenza -associated deaths	95% CI	All causes influenza -associated deaths rates per 100,000 population	95% CI
1980-1981	-	12-3	4	A(H3N2)	5,638	5,044 - 6,232	39.1	34.7 – 43.6
1981-1982	-	1-2	2	B	0	-	0.0	-
1983-1984	-	3-4	1	A(H1N1)	2,487	2,053 - 2,901	15.7	12.7 – 18.8
1984-1985	-	1-2	2	A(H3N2)	1,802	1,468 - 2,136	12.0	9.7 – 14.4
1985-1986	-	12-2	3	A(H3N2)	4,784	4,193 - 5,375	28.4	24.6 – 32.1
1986-1987	-	1	1	A(H1N1)	1,202	861 - 1,543	6.7	4.7 – 8.7
1987-1988	-	1	1	B, A(H1N1)	0	-	0.0	-
1988-1989	-	12-1	2	A(H1N1), A(H3N2)	2,530	2,053 - 3,007	13.7	11.0 – 16.4
1989-1990	-	1-2	2	A(H3N2)	3,920	3,516 - 4,324	19.7	17.6 – 21.6
1990-1991	-	12-2	3	B	2,781	2,313 - 3,249	13.7	11.4 – 16.0
1991-1992	11-1	12-2	3	A(H3N2)	2,845	2,466 - 3,244	14.1	12.1 – 16.1
1992-1993	2-4	3	1	B	107	17 - 197	0.6	0.1 – 1.2
1993-1994	11-1	12-1	2	A(H3N2)	3,529	3,601 - 3,997	15.9	13.8 – 18.0
1994-1995	1-2	11	1	B	0	-	0.0	-
1995-1996	10-1	11-12	2	A(H3N2)	1,892	1,527 - 2,257	8.8	7.0 – 10.7
1996-1997	11-2	12-3	4	A(H3N2)	5,533	4,997 - 6,069	25.5	22.8 – 28.3
1997-1998	x	3-4	2	A(H3N2)	308	91 - 525	1.1	0.3 – 1.9
1998-1999	12-2	12-4	5	A(H3N2)	8,514	7,908 - 9,120	36.1	33.2 – 39.0
1999-2000	1-2	1-2	2	A(H3N2)	3,363	2,904 - 3,822	14.2	12.1 – 16.2
2000-2001	x	x	0	B	0	-	0.0	-
2001-2002	1-2	1-3	3	A(H3N2)	2,145	1,722 - 2,568	8.8	6.9 – 10.8
2002-2003	11-12	11-12*	2	B	0	-	0.0	-
2003-2004	10-12	10-12*	2	A(H3N2)	950	656 - 1,244	3.1	2.1 – 4.1

Table 3.1: Characterization of influenza seasons from 1980-1981 to 2003-2004 according to the duration of the epidemic periods, dominant (sub)type of influenza virus, all causes influenza associated excess absolute deaths and age-standardized death rates. ISS: Influenza surveillance system, ISM: Influenza specific mortality. * Information is based on ILI surveillance and influenza virus activity; x - no epidemic period detected; NA, data not available.; Month numbers 1-January to 12-December ** Information on the season dominant type of virus for seasons 1982-83 to 1989-90 was obtained from the WHO. From 1990-91 to 2004-05 this information was obtained by the Portuguese ISS.

	All Causes		Cerebrovascular diseases		Ischemic heart disease		Diseases of the respiratory system		Pneumonia and Influenza		Chronic respiratory diseases	
	Rate	%IS	Rate	%IS	Rate	%IS	Rate	%IS	Rate	%IS	Rate	%IS
All individuals	12.97	3.0	2.88	3.2	0.69	2.0	3.14	9.3	1.45	9.90	0.77	6.9
Individuals ≥ 65	155.75	4.00	38.48	3.50	8.19	2.30	35.69	10.1	16.86	11.6	9.09	10.00
Age group												
0-4	2.57	1.3	***	***	***	***	0.40	2.5	0.7	6.0	***	***
5-54	0.97	0.80	0.08	1.20	0.07	1.10	0.4	8.1	0.20	7.2	0.1	4.1
55-64	14.2	2.00	1.02	1.10	0.93	1.20	4.0	9.8	1.0	7.7	1.0	5.3
65-69	34.96	2.60	4.51	1.7	3.49	2.20	11.50	11.80	3.5	12.0	5.0	10.7
70-74	78.80	3.4	19.19	3.3	3.5	1.3	18.9	10.0	6.2	10.3	5.6	6.5
75-79	170.7	4.2	41.85	3.5	10.4	2.5	37.2	10.4	14.8	11.3	10.7	7.3
80-84	332.3	4.6	96.50	4.3	17.4	2.8	74.7	11.2	40.9	14.3	14.4	6.3
≥ 85	825.8	5.5	207.10	4.8	33.5	3.2	174.2	11.8	103.5	14.1	33.6	8.9
Dominant virus*												
A(H3)	18.0	4.2	3.8	4.3	1.0	2.8	4.5	13.4	2.1	13.9	1.1	9.9
A(H1) or B	4.1	0.9	1.2	1.3	0.1	0.4	0.7	2.1	0.4	2.9	0.2	1.7
Ratio A(H3)/A(H1) or B	4.4	4.5	3.1	3.4	6.8	7.5	6.8	6.56	5.5	4.8	5.0	5.7
p**	0.001	0.001	0.015	0.004	0.001	0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.001	< 0.001

Table 3.2: Average rates of excess mortality associated with influenza epidemics and proportion of deaths attributable to influenza by disease outcome, age group, and dominant viral subtype, Portugal 1980-2004. Rates are per 100,000 population. * age-standardized rates; % IS: proportion of winter death attributable to influenza; calculated as the ratio of excess deaths to death occurring from October to May, by age group, mortality outcome, and season; ** Mann-Whitney test for comparison of excess mortality during A(H3) and A(H1) or B seasons; *** data not presented due to small death counts

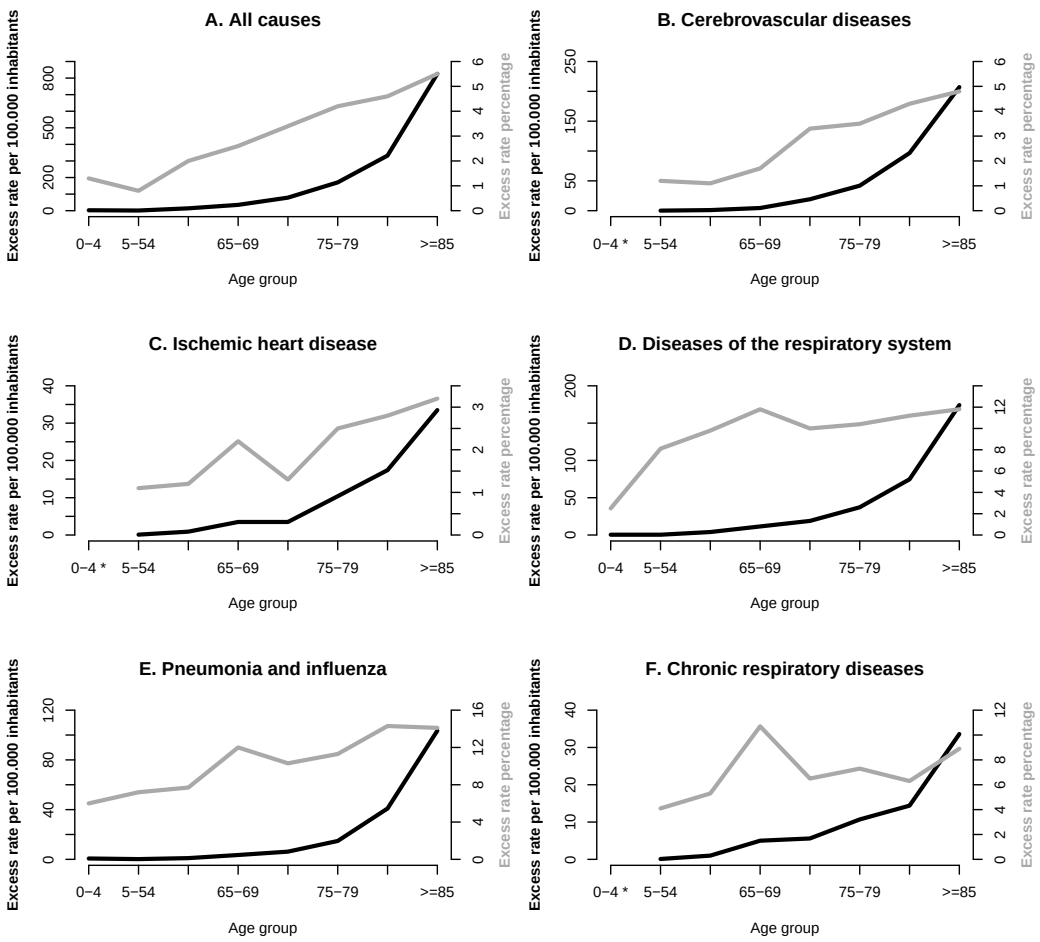

Figure 3.2: Age-specific influenza excess mortality burden. Average rates (per 100,000 persons) and proportion of winter mortality associated with influenza epidemics from 1980-1981 to 2003-2004 by age group: A. All causes, B. Cerebrovascular diseases, C. Ischemic heart diseases, D. Diseases of the respiratory system, E. Pneumonia and Influenza, F. Chronic respiratory diseases.(* data not presented due to low annual number of deaths). The proportion of winter mortality attributable to influenza was calculated as the ratio of seasonal excess mortality to mortality occurring during Oct to Mar, for each disease outcome and age group.

3.5.4 Comparison of influenza-related excess mortality and morbidity

We compared seasonal excess mortality patterns in seniors over 65 years old (which account for the majority of influenza-related deaths) with morbidity patterns in ILI consultation rates in the same age group. The correlation between age-standardized influenza-associated mortality rates for the studied causes and seasonal ILI consultation rates ranged between 0.64 and 0.83 ($p < 0.05$; Table 3.3). These results show that high ILI seasons were associated high excess mortality seasons among seniors. In Figure 3.3 we present the results for P&I.

	All causes	CVD	IHD	DRS	PI	CRD	ILI(1)	Injuries
All causes	1	0.950*	0.845*	0.926*	0.950*	0.824*	0.765*	0.026
CVD		1	0.783*	0.786*	0.857*	0.652*	0.641*	-0.082
IHD			1	0.855*	0.828*	0.806*	0.829*	0.094
DRS				1	0.952*	0.949*	0.807*	0.079
PI					1	0.838*	0.743*	-0.0002
CRD						1	0.794*	0.151
ILI(1)							1	0.426

Table 3.3: Correlation matrix between seasonal age-standardized excess rates. Injuries are used as a control time series which should not be associated with influenza virus circulation. CVD: cardiovascular disease; IHD: ischemic heart disease; DRS: diseases of the respiratory system: PI: Pneumonia and Influenza; CRD chronic respiratory disease: * $p < 0.05$; (1) correlation with ILI was only performed for the group of 65 and plus years of age.

3.5.5 Influenza epidemic periods validation and model diagnostics

We used two approaches to define epidemic periods; one was based on influenza-specific deaths and the other on the Portuguese ISS, combining ILI and influenza laboratory surveillance. The two approaches proved to be consistent in the period when both datasets overlapped, 1991-2002 (Table 3.1). There was a lag of 0 to 1 months between epidemic periods defined by the Surveillance System and influenza-specific mortality, except for the mild 1994-1995 season, where epidemic periods did not match.

Model fit was relatively good as there was no auto-correlation in the residuals of the SARIMA models in 46 out of the 48 models; however, normality of the residuals was always rejected (See details in Appendix A). Seasonal excess mortality estimates were generally consistent across the various causes of death studied (correlation range,

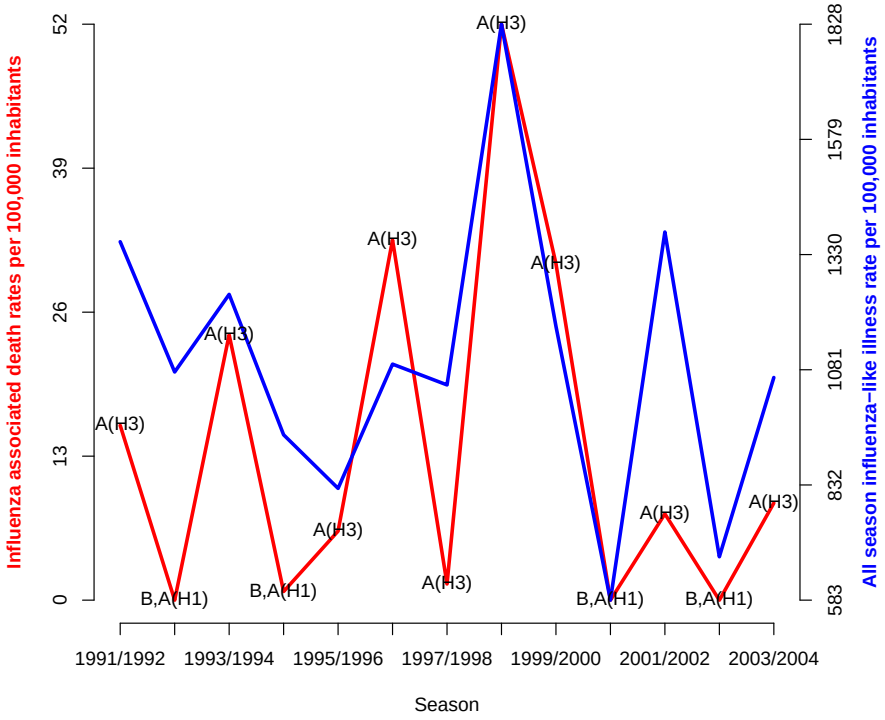

Figure 3.3: Seasonal rates of excess pneumonia and influenza and seasonal rates of influenza like illnesses in the elderly population over 65 years, showing dominant circulation strains of virus.

0.65 to 0.95, $p < 0.05$; Table 3.3). In particular the correlation between all-causes excess deaths and excess P&I deaths was very high 0.95 ($p < 0.001$).

3.5.6 Specificity analysis

Next, we applied the same excess mortality approach to deaths from injuries, to check that the method did not spuriously attribute deaths to influenza in unrelated outcomes. The seasonality of injuries was out-of-phase with that of influenza and displayed summer peaks (Figure 3.1). On average for the 24 seasons under study, our excess mortality approach attributed only 0.66% of all injury deaths to influenza (See details in Appendix B), which is much lower than our estimates for other studied causes of death (2-10%, Table 3.2).

In addition, seasonal age-standardized excess mortality rates from injuries were

not correlated with excess rates from causes that are traditionally associated with influenza ($p > 0.05$ for correlation with excess deaths from P&I, chronic respiratory diseases; all respiratory diseases; cardiovascular disease; ischemic heart disease; all-causes; Table 3.3). Finally, there was no correlation between seasonal excess deaths from injuries and seasonal ILI attack rates in seniors.

3.6 Discussion

This is the first study to provide estimates of influenza seasonal mortality burden for the Portuguese population in a comprehensive way, which is an important step for designing rational national public health measures and in comparing rates with other countries. On average, 2,475 excess all-cause deaths were attributable to influenza in Portugal each winter during 1980-2004 (range=0 to 8,514), corresponding to a seasonal excess mortality rate of 24.7 per 100,000, and a 2000 world population age-adjusted rate of 13 per 100,000.

Our estimates of influenza-related excess mortality rate are in accordance with previous studies. In the US, Thompson et al. estimated an average excess mortality rate of 19.6 per 100,000 inhabitants [33], while Schanzer et al.[58] calculated that influenza was responsible for an excess mortality rate of 13.0 per 100,000 in Canada. In Europe, excess mortality rates varied between 16.0 per 100,000 in Germany [17] and 26.0 per 100,000 in the Czech Republic [54].

Despite the fact that P&I and all-cause excess mortality are considered reliable indicators of influenza mortality impact [33, 15], here we used a broader approach and evaluated the impact of influenza epidemics on various causes of mortality in Portugal [49]. Although excess mortality was observed for 7 causes of death that are traditionally linked to influenza, in this study, the highest proportion of October-May deaths attributable to influenza was observed for respiratory-related illnesses, with proportions ranging between 6.9 and 9.9%. Since influenza can precipitate other severe respiratory conditions, a higher influenza attributable proportion is expected for respiratory diseases, especially for pneumonia [65]. Similar results were obtained in Canada [58] and the US [16, 33], where pneumonia had the highest estimated percentage (8%) of deaths attributable to influenza. By contrast, all-cause mortality is a less specific outcome and we estimate that only 3% of total October-May deaths in Portugal can be attributable to influenza (4% in seniors over 65 years). This compares

with estimates of 5% in the US [16] and 4.8% in Italy [18]. We note there is substantial inter-annual variation in this percentage in Portugal and in other countries [16, 18], as the burden of influenza epidemics varies considerably depending on circulating strains and population immunity.

It is interesting to note that P&I excess deaths only captured on average 11.2% of the total mortality burden of influenza in Portugal. By contrast, excess cerebrovascular mortality was the major contributor to total influenza-associated deaths (22.3%). This is not entirely unexpected, as cerebrovascular disease is the primary cause of death in Portugal [66]. For comparison, estimates obtained in Canada show that the proportion of excess deaths captured by P&I was 22.7% (8% for influenza and 14.7% for pneumonia), while cerebrovascular disease only contributed to 6.5% of influenza-attributable deaths [58]. The major contributor to influenza excess mortality in Canada was P&I and ischemic heart disease, the latter outcome representing 22.9% of all influenza-associated deaths. This finding is particularly relevant because ischemic heart disease is the leading cause of death in Canada [67]. Whether between-country differences in leading causes of deaths and their association with influenza result from differences in coding practices or health status remains unclear and warrants further studies.

Overall, the selection of the most appropriate cause of death to adequately measure the burden of influenza remains a point of debate. On the one hand, the usage of influenza-specific deaths grossly underestimates the actual impact of the influenza virus [28]. For example, in this study, the ratio of all influenza-attributable deaths to deaths coded specifically as influenza reached 23, and the number of influenza-coded deaths declined dramatically after 2002, in parallel with the transition of WHO International Classification of Diseases from revision 9 to 10. On the other hand, all-cause mortality lacks specificity and provides less accurate estimates, especially for milder seasons [33]. As reviewed in [49] and [68], the causal relationship between influenza and cardiovascular disease is well established, so that mortality from cardiovascular diseases should be more systematically included in influenza mortality studies.

Seasons with highest estimates of excess deaths were characterized by a dominance of the A(H3) influenza viral subtype in Portugal, a pattern that was consistent across all mortality outcomes. A higher level of hospitalizations and deaths associated with the A(H3) subtype is already well documented [16, 33, 17]. In our study, influenza B seasons had generally very low impact. In particular, the 5 seasons associated with no

excess mortality were dominated by influenza B or had high proportion of B viruses. Influenza B is generally more prevalent in children and young adults, who have lower risk of severe influenza-related outcomes than seniors.

The age-standardized ratio of A(H3) vs (A(H1) or B) excess mortality was respectively 4.8 and 3.6 for P&I and all causes in Portuguese seniors. Similar results have been obtained in Italy for P&I (4.5) [18]. By contrast, the subtype rate ratio for all-cause mortality was much lower, in the US (2.8) and in Italy (2.9), after taking into account population age differences [16, 18]. These divergences could be explained by differences in the strength and sampling protocols of laboratory virus surveillance systems in each country and/or the definition of subtype-dominant seasons.

On average, 90% of influenza-associated excess deaths occurred in the 65 years age group. Studies in the Netherlands [32] and the US [33] estimated that 90-95% of influenza-related deaths occurred in seniors despite the use of a different methodological approach and time periods. In Portugal, the unadjusted rate of all-cause excess mortality was 165 per 100,000 on average in people 65 years. This value was 12-65% higher than estimates for Italy [18], Netherlands [32], US [16], and Canada [58]. One possible explanation for these differences could be the lower level of influenza vaccine coverage in Portuguese seniors compared to these countries. From 1998/99 to 2003/04, the influenza vaccine coverage in the elderly increased from 31% to 47% in Portugal [69]. While this progress is significant, these values were considerably lower than coverage estimates in the other countries over comparable time periods. For instance, seasonal influenza vaccine coverage increased from 32% in 1993/94 to 61% in 2000/2001 in Italy [18] and from 31% in 1988/89 to 65% in 2000/2001 in the US [16]. We note that no decline in influenza attributable excess mortality was observed in Italy and the US as vaccination coverage increased [16, 18], suggesting that vaccination is not the only explanation for the higher excess mortality rates observed among Portuguese seniors.

The average excess mortality rate for the children aged 0-4 years old in this study was 2.6/100,000, which is an order of magnitude higher than findings from previous studies. A recent study in the Netherlands did not find any excess mortality associated with influenza in the younger groups during 1997-2003 [70]. Another study in the US showed that the average influenza-associated excess death rates for children < 1 years and 1-4 years of age were 0.3 and 0.2 per 100,000, respectively [33]. The US study also estimated a much higher rate of excess mortality due to respiratory syncytial virus

(RSV) in these age groups, a finding that is in line with many other studies that establish RSV as a predominant respiratory virus in young children [71, 72, 73]. It is possible that our estimates of influenza excess mortality for young Portuguese children could be confounded by the co-circulation of RSV. Unfortunately, no information on the circulation of RSV is currently available for Portugal.

In this study, we found that excess mortality increased exponentially with age after 50 years of age, while the proportion of October-May deaths attributable to influenza increased more linearly with age. This indicates that the increased risk of influenza-related mortality with age is driven by the increase in background risk of death and may even suggest that the etiological fraction of influenza does not increase with age as much as other potential causes.

There are some caveats worth noting in our study. First, we relied on statistical models based on seasonal linear regression and SARIMA to estimate by an ecological method the burden of influenza. We have compared estimates from our SARIMA approach with those from a traditional Serfling seasonal linear regression [16, 48] and found very consistent results. In addition, we report a strong association between seasonal rates of influenza like illness and excess mortality for causes that are traditionally linked to influenza in the elderly population. This result confirms the robustness of our estimates of influenza-attributable excess death, since independent indicators of influenza-related excess mortality and morbidity coincide. We have also checked the specificity of our approach by estimating excess deaths attributable to influenza in injury mortality time series. This specificity analysis revealed that the proportion of injury deaths "attributable" to influenza was very low, and that season-to-season variations in excess injuries were not associated with ILI activity. This supports the conclusion that our approach provides excess mortality estimates that are specific to influenza.

Chapter 4

Nowcasting influenza epidemics using non homogenous hidden Markov models[1]

4.1 Introduction

During public health threats like infectious diseases outbreaks (SARS, West Nile virus, pandemic influenza, etc) knowing in real-time the epidemic trend, spatial distribution and impact in terms of medical consultations, hospitalizations and deaths is essential to identify the most appropriate measures to control the disease spread and mitigate its effect [74, 75, 76]. It is widely accepted that the most proper tool to acquire reliable information on disease spread and impact remains in the use of stable disease surveillance systems defined as *"...the ongoing, systematic collection, analysis, interpretation and dissemination of disease related data for public health action..."* [74].

As mentioned before, the epidemiological surveillance of influenza activity in Europe is supported by sentinel systems, formed by GP's, that provide weekly the number of consultations with patients presenting ILI symptoms. These figures divided by the number of patients in the GP list or the total number of consultations on that week enable the estimation of ILI incidence rates. On other hand GP sentinel networks also provide, for some of the ILI patient reported, nasopharyngeal swabs that after laboratory confirmation support the virological surveillance of the influenza

[1]This Chapter was based on the paper Nunes B, Natário I, Carvalho L. Nowcasting influenza epidemics using non homogenous hidden Markov models. submitted for publication in the peered reviewed journal *Statistics in Medicine* in July 2011.

activity [6, 77].

At the European level the ISS is managed by the ECDC through a network of countries named European Influenza Surveillance Network. In this framework each Wednesday all participating countries upload age-specific ILI rates and virological information corresponding to the previous week[2] in a web-based system (TESSy: The European Surveillance System). The European influenza activity report gathering information from all the participating countries (WISO: weekly influenza surveillance overview) is issued by the ECDC on the following Friday [6].

In a no delay reporting situation one could consider that by Monday all the information to estimate ILI rates for the previous week should be complete. So for a particular week t, ILI rate estimated on Monday of week $t+1$ could be considered the zero day delay ILI rate estimate. Having made these considerations, in the present European surveillance system, countries upload age-specific ILI rates with a 2 days delay and ECDC issues the WISO with a 4 days delay.

Timeliness of a public health surveillance system is one of its most important characteristics, given that it is crucial for its capacity of a timely intervention [74, 75]. For this study, timeliness will be considered as the time elapsed from the disease onset to the generation of an automated alert.

The prospective detection of the beginning of the epidemic period has been done by a variety of statistical methods such as regression techniques, time series methods, methods of statistical process control and also on statistical multivariate analysis using multiple data sources. A comprehensive review of these methods can be found in [78, 79, 80, 81]. Some of these methods were implemented in the R package `surveillance` [82] and can be easily used for ongoing disease surveillance. Nevertheless, for all these methods the main goal is to identify the first week of the epidemic period, when the ILI incidence rate indicates levels of influenza activity that can be classified as epidemic.

Regarding the ISS based on GP sentinel networks, all these methods were only applied to data referring to the previous week. This means that the online detection algorithms could only report an alert at most by the beginning of the following week. Despite that, data providers (GP) send the data to the surveillance system on a daily basis. This has became more promptly since some surveillance systems use web-based systems where GP can enter data during consultation with the patient

[2]ISO definition: Monday through Sunday

[83] or use computer routines to capture data from GP electronic medical records [84, 85]. Therefore this daily data stream could provide the surveillance systems sufficient information to assess the current situation, without time delay, enabling the real-time early detection of the epidemic start, peak and end.

The process of predicting the present situation using the available incomplete information has been considered of high interest by public health officials, mainly during the pandemic (H1N1)2009 receiving the term of *nowcasting* [86]. The use of incomplete information was already applied to nowcast on a weekly basis, the number of influenza A(H1N1)2009 hospitalizations during the 2009-10 pandemic in the Netherlands with considerable success [12].

In this work we developed a statistic model that on a weekly basis, uses all data collected by a surveillance system before the end of the week, to nowcast two measures of interest: the ILI incidence rate and the influenza activity state, epidemic or non-epidemic, that will be reported by Wednesday of the following week to ECDC. Given this objective, we were able to show the adding value of using a non homogenous HMM, being the advantage mainly shown by its ability of using covariates, with early information on the epidemic (like weekly ILI cases laboratory confirmed, early estimates of the ILI rate, ILI rate for other age groups, etc), to model time varying influenza activity state transition probabilities, in opposition to the homogenous model used in prior studies [23, 24, 25], where the state transition probabilities remain constant overtime. Additionally we were also capable of introducing covariates in the response variable (Wednesday ILI rate) in order to nowcast its specific value. To our knowledge this work represents the first attempt to use non homogenous HMMs in a disease surveillance problem with the objective of early detect an outbreak and nowcast its evolution.

4.2 Hidden Markov models

4.2.1 Model specification

Generally a HMM assumes that an observed time series, y_t with $t = 1, 2, ...,$ is a realization of a stochastic process $\{Y_t : t = 1, 2, ...\}$, where the distribution of each Y_t is conditionally determined by an unobserved (hidden) discrete Markov process $\{S_t : t = 1, 2, ...\}$, taking values in a m-states set $\{1, 2, ..., m\}$.

This unobserved stochastic process is assumed to be, in the most usual form, an

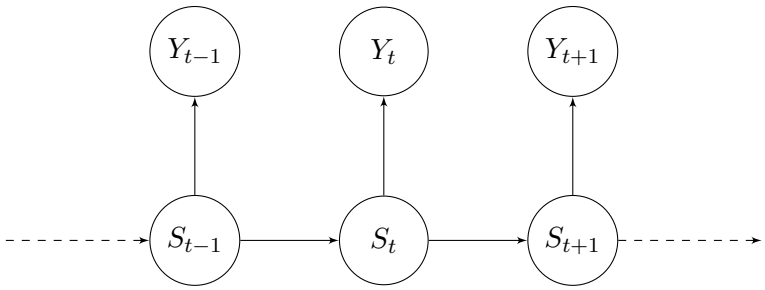

Figure 4.1: Direct graph of an order one HMM

order one homogenous Markov chain, with initial distribution $P(S_1 = i)$ and state transition probabilities given by $\gamma_{j,i} = P(S_t = i|S_{t-1} = j)$ for any $i, j \in \{1, ..., m\}$, the elements of the transition probabilities matrix $\Gamma = (\gamma_{j,i})$.

In this model Y_t is considered to be a state dependent process, which means that when S_t is known, the distribution of Y_t is determined by its value. As a consequence Y_t conditionally on S_t is independent from its previous observations and previous hidden states (Figure 4.1). More specifically the state dependent distributions are:

$$f_i(Y_t = y|S_t = i, \boldsymbol{\theta}_i)$$

with $i \in \{1, ..., m\}$ and $t = 1, 2, ...$, where f_i represents the state i specific density (or mass) function of Y_t with parameters $\boldsymbol{\theta}_i$.

Given this considerations the likelihood of an observed time series $y^T = (y_1, ..., y_T)$ assumed to be generated by a m-state HMM with parameters $\Psi = (\boldsymbol{\theta}, \Gamma)$, can be expressed as the sum for all possible hidden state sequences s^T:

$$f(y^T|\boldsymbol{\Psi}) = \sum_{s^T} f(y^T|s^T, \boldsymbol{\theta}) f(s^T|\Gamma)$$

where

$$f(y^T|s^T, \boldsymbol{\theta}) = \prod_{t=1}^{T} f(y_t|s_t, \boldsymbol{\theta}_{s_t})$$

and

$$f(s^T|\Gamma) = P(S_1 = i) \prod_{t=2}^{T} \gamma_{s_{t-1}, s_t}.$$

This family of models have been useful in many problems some examples are: the analysis of DNA sequences [87], machine speech recognition [88], longitudinal studies on epileptic seizures [89], misclassification of diseases outcomes [90], hospital infections [91] and, of course, disease surveillance [23, 24, 25, 92].

4.2.2 Application to influenza surveillance

The HMM were first applied to ILI surveillance data in 1999 by Le Strat and Carrat [23]. In their work they proposed a two states homogenous HMM, epidemic and non-epidemic, where in each of these states the weekly ILI rate was described by a normally distributed cyclical model of period 52 plus a linear trend. Following this first work others have emerged using the same line of approach, from which we point out the work of Rath et al 2003 [24], that proposed independent and identical distributed (i.i.d) exponential distributions for rates in the non-epidemic state and i.i.d normal distributions for the rates in the epidemic state, and the work of Martinez-Beneito et al 2008 [25], that modeled the one week lag differences of the weekly ILI rates, considering that those for the non-epidemic state were normally distributed with mean zero, and those for the epidemic state were described by an order one autoregressive model.

As stated before the application of these models to ILI data had always the goal to classify weeks as epidemic or non epidemic using all available data [23, 24] or, like in Martinez-Beneito et al 2008 [25], to do this on an online basis, classifying the last week available as epidemic or non-epidemic. This last paper has used a bayesian approach for the hidden states and parameters estimation.

It is important to notice that these approaches did not have the objective of predicting or forecasting the ILI rates or even to forecast the hidden Markov chain state for the forthcoming weeks.

4.2.3 The non-homogenous HMM

Given our objective of nowcasting the current week ILI rate (not yet observed) and the respective influenza activity state using the incomplete data collected by the ISS, we need a model that enables the use of covariates with forecasting capacity. For this purpose we propose the use of a HMM that allows the introduction of covariates to model the weekly ILI rate and the sate transition probabilities. This late innovation implies the use of a non homogenous HMM instead of a homogenous one as in the previous literature [23, 24, 25], where the transition probabilities from the non-epidemic to the epidemic state (the epidemic beginning) and from the epidemic to the non-epidemic state (the end of the epidemic) were invariant in time.

Applications of non homogenous HMM have been described in a large variety of

subjects, like modeling the duration of business cycles [93], precipitation occurrence [94], the association between sulphur dioxide and meteorologic variables [95], estimation of infection and recovery rates for highly polymorphic parasites [96]. More specifically, we propose the application of a HMM where the state transition probabilities are given by a time-dependent matrix Γ^t, with elements $\gamma^t_{j,i} = P(S_t = i|S_{t-1} = j)$ for any $i, j \in \{0, 1\}$ and $t = 2, ..., T$, where 0 and 1 represent, respectively, the non-epidemic and the epidemic state of influenza activity,

$$\Gamma^t = \begin{bmatrix} \gamma^t_{0,0} & \gamma^t_{0,1} \\ \gamma^t_{1,0} & \gamma^t_{1,1} \end{bmatrix}.$$

Choosing the transition probabilities to be modeled by a logistic function of the covariates [97, 95],

$$\text{logit}(\gamma^t_{j,i}) = \ln\left(\frac{\gamma^t_{j,i}}{\gamma^t_{j,j}}\right) = \boldsymbol{\alpha}_{j,i}\boldsymbol{Z}_t$$

$$\gamma^t_{j,i} = \frac{\exp(\boldsymbol{\alpha}_{j,i}\boldsymbol{Z}_t)}{(1 + \exp(\boldsymbol{\alpha}_{j,i}\boldsymbol{Z}_t))}$$

for any $i, j \in \{0, 1\}$ and $j \neq i$, where $\boldsymbol{Z}_t = (1, z_{1,t}, ..., z_{q,t})$ is a vector of q covariates measured in time t. The non homogeneity of the state transition probabilities can preclude the stationarity of the Markov chain [97]. This can happen mainly when the covariates used are functions of time. In this situation, an initial distribution for the non homogenous hidden Markov chain is defined by $\boldsymbol{\delta} = (\delta_0, \delta_1)$, $\delta_i = P(S_1 = i)$ with $i \in \{0, 1\}$.

To better understand the advantages of the non homogenous HMM we apply this model to the data collected by the Portuguese ISS. For comparison purposes we also apply to the same data an equivalent model with a homogenous hidden Markov chain.

4.3 Data description

In Portugal the clinical component of the ISS is assured by a network of 144 voluntary GP's ("Medicos Sentinela") that, since 1989, weekly report the ILI cases[3] that occurred among their patient list [83, 98] to the surveillance system hub (since year 2000, the Department of Epidemiology of Instituto Nacional de Saúde Dr. Ricardo Jorge), by regular mail or using a web-based questionnaire. The population under

[3]ILI definition used by the ICPC - WONCA

observation, comprising the total of the GP patients lists, represents a sample fraction of 2.5% of the total population.

The laboratorial component is coordinated by the National Influenza Reference Laboratory and consists in receiving swabs from ILI cases sent by a fraction of GPs from the sentinel system and from the network of emergency departments from hospitals and health centers for virological surveillance purposes.

The Portuguese system, as part of the EISN, calculates and reports each Wednesday to TESSy system, the age-specific ILI incidence rates and the number of ILI cases with laboratory confirmed influenza infection corresponding to the week before[4]. This is done because there are delays in the system, namely the time between the patient onset of symptoms to the GP visit and the GP reporting time to the surveillance system.

In order to have a predictor of the ILI rate reported to ECDC by Wednesday, the Portuguese ISS has been calculating the age-specific ILI incidence rates for each week with the incomplete data gathered on Friday of that same week, from season 2008-2009 to 2010-2011. Our main goal is to access if there is enough information in this variable to be used as predictor for an early detection of the epidemic start, peak and end.

In synthesis, the data used in this study is the weekly ILI incidence rate (per 100,000 inhabitants) of week t calculated by Friday of week t, referred as $y_{t(t)}$, and the weekly ILI rate calculated by Wednesday of week $t+1$, referred as $y_{t(t+1)}$ (Figure 4.2), for the period from week 40 of 2008 to week 16 of 2011. Virological data consists of the weekly number of ILI cases with laboratory confirmation for the influenza virus corresponding to week t, but calculated by Wednesday of week $t+1$: $v_{t(t+1)}$.

4.4 Models

The main objective is to nowcast the influenza activity state and the Wednesday ILI rates for week t, calculated in week $t+1$, using available data from covariates at time t, namely the early estimate of the ILI rates by Friday and the number of ILI cases tested positive for influenza for the previous week, estimated by Wednesday of the present week, $v_{t-1(t)}$. To achieve this goal a HMM, based on the work developed by Paroli and Spezia 2008 [95], that allows the inclusion of covariates in response

[4]The date of reference is the disease onset date

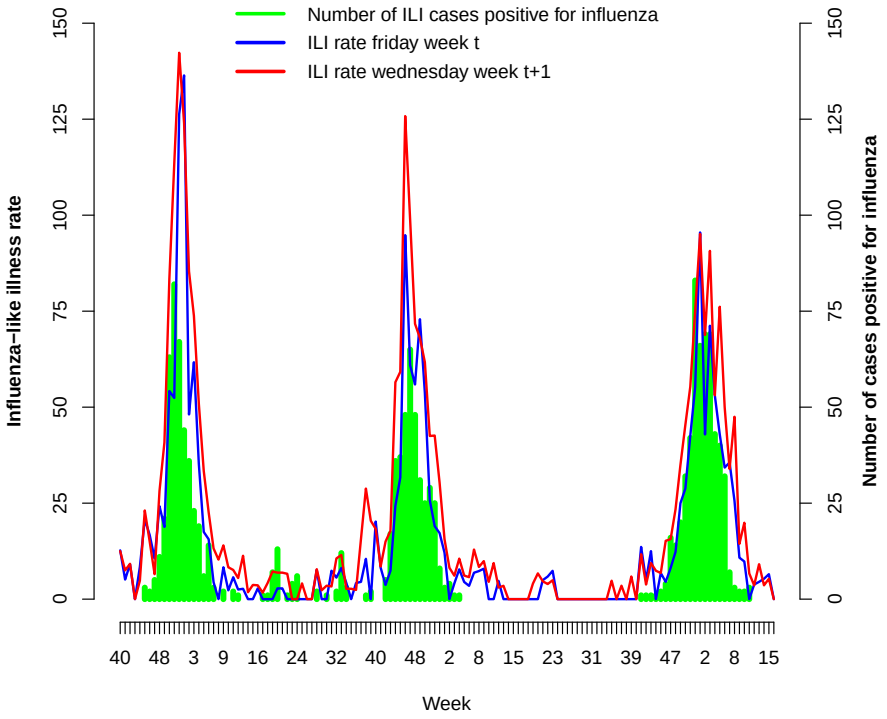

Figure 4.2: Influenza-like illness incidence rates calculated by Friday of week t and by Wednesday of week $t+1$.

variable and in the transition state probabilities is proposed.

From our observation, the influenza epidemic is sustained when the number of ILI cases tested positive for influenza in a certain week is high, e.g 20. On the other hand if the number ILI case confirmed for influenza is zero, the influenza activity state is clearly non epidemic. Based on this, if an increase in the ILI rate is observed without ILI cases confirmed for influenza, one can not assume that this increase is due to an influenza epidemic, but can be related to the circulation of other respiratory viruses. Using this empirical thresholds, two covariates that are functions of the early estimate of the ILI rate and of the number of ILI cases positive for influenza in the previous week are proposed. The objective of these two covariates is to enhance the capacity of the models to discriminate an observed ILI rate as belonging to either the epidemic or non epidemic period, keeping its prediction ability: one covariate to

model the response variable in the non epidemic period,

$$y_{t(t)0} = \begin{cases} y_{t(t)} & v_{t-1(t)} \leq \nu_0 \\ 0 & \text{otherwise} \end{cases}, \tag{4.1}$$

and the other to model the response variable in the epidemic period,

$$y_{t(t)1} = \begin{cases} y_{t(t)} & v_{t-1(t)} \geq \nu_1 \\ 0 & \text{otherwise} \end{cases}. \tag{4.2}$$

In this study ν_0 was set at 20 and ν_1 to 1. As can be seen, both variables share a common part of the early estimate of the ILI rate $y_{t(t)}$, i.e. the weeks where the number of ILI cases tested positive, in the previous week, is higher or equal to 1 and lower or equal to 20.

Baring this in mind, the weekly incidence rate calculated by Wednesday $y_{t(t+1)}$ is assumed to be described by the following equations, conditional on the influenza activity state:

$$y_{t(t+1)} = \begin{cases} \mu + \beta_1 \cos(\frac{2\pi t}{52}) + \beta_2 \sin(\frac{2\pi t}{52}) + \theta_{0,1} y_{t(t)0} + e_{t,0} & S_t = 0 \\ \mu + \beta_1 \cos(\frac{2\pi t}{52}) + \beta_2 \sin(\frac{2\pi t}{52}) + \theta_{1,1} y_{t(t)1} + \theta_{1,2} y_{t(t)1}^2 + e_{t,1} & S_t = 1 \end{cases} \tag{4.3}$$

were $e_{t,i} \sim N(0, \tau_i)$, with precision $\tau_0 > \tau_1$, $t = 1, ..., T$ and $i \in \{0, 1\}$. Given that the epidemic state is characterized by values of ILI rate higher than the ones in the non epidemic state, a constraint that forces the variance of the epidemic state to be higher than the variance of the non epidemic state is included in the model.

In this model we propose a common component for the epidemic and non-epidemic period that reflects the baseline behavior of the ILI rates. This component is constituted by a common mean μ and a cyclical component of period 52 weeks. The difference between the equations for each state is set on the association with the predictor, i.e. the state specific variables that are functions of the early estimate of the ILI rate on Friday of week t, presented in equations 4.1 and 4.2. So, for the non-epidemic model a linear association with $y_{t(t)0}$ is considered and, on other hand, the epidemic period is described by a quadratic association with $y_{t(t)1}$. We propose these two state dependent equations, from the observation of Figure 4.3 and with the rational that, during the non epidemic period, i.e. for weeks with zero or a very small number of ILI cases tested positive, the number of ILI cases are uniformly distributed within the week, so the relation between the early estimate of ILI rate by Friday and

the ILI rate estimated by Wednesday of the following week is linear. On the other hand, during the epidemic period, the distribution of the ILI cases within each week will not be uniform, presenting a sharp growth or decrease, respectively, in the beginning and at the end of the epidemic. Given this, we propose for the epidemic period, the quadratic form to model the association between the two ILI rates estimates.

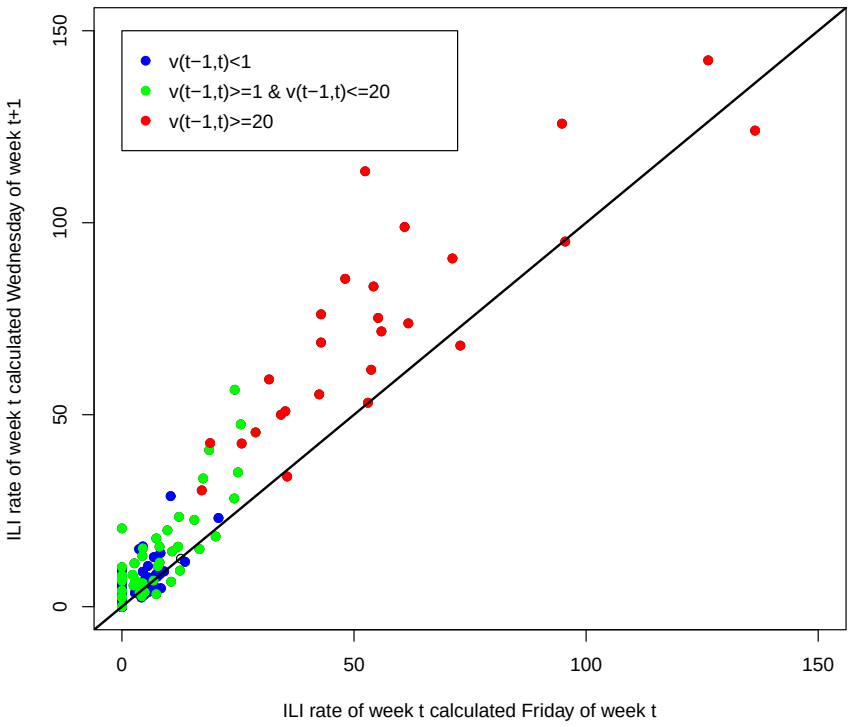

Figure 4.3: Association between ILI incidence rates calculated by Friday of week t and by Wednesday of week $t+1$ according to the number of ILI cases tested positive for influenza in the previous week $v_{t-1(t)}$. Black line represents $y_{t(t+1)} = y_{t(t)}$

For the hidden Markov chain state transition probabilities a time-varying matrix with elements $\gamma_{j,i}^t$ for any $i, j \in \{0, 1\}$ and $t = 2, ..., T$ was considered. For a specific week t, $\gamma_{0,1}^t$ and $\gamma_{1,1}^t$ represent the probability that in week t the influenza activity is epidemic given that in the week before the influenza activity was respectively in the non-epidemic or in the epidemic state.

Given these considerations, three models are proposed (Models 0, 1 and 2). All three share the same model for the response variable, expressed in equation 4.3, but

have different choices for the state transition probabilities. Models 1 and 2 have a non-homogenous hidden Markov chain, that are differentiated according to the covariates used in the logistic function that models the transition probabilities:

- Model 1:

$$\text{logit}(\gamma_{j,i}^t) = \ln\left(\frac{\gamma_{j,i}^t}{\gamma_{j,j}^t}\right) = \alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)}$$

$$\gamma_{j,i}^t = \frac{\exp(\alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)})}{(1 + \exp(\alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)}))}$$

- Model 2:

$$\text{logit}(\gamma_{j,i}^t) = \ln\left(\frac{\gamma_{j,i}^t}{\gamma_{j,j}^t}\right) = \alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)} + \alpha_{j,i,2} v_{t-1(t)}$$

$$\gamma_{j,i}^t = \frac{\exp(\alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)} + \alpha_{j,i,2} v_{t-1(t)})}{(1 + \exp(\alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)} + \alpha_{j,i,2} v_{t-1(t)}))}$$

for any $j, i \in \{0, 1\}$ and $j \neq i$.

More specifically, for Model 1, the transition probabilities from the non-epidemic state to the epidemic state and *vice versa* are modeled by the early estimate of the ILI incidence rate. In Model 2 transition probabilities are modeled by the early estimate of ILI incidence rate and by the absolute number of ILI cases tested positive for influenza in the week before.

Finally for Model 0 a homogenous hidden Markov chain is used, where $\gamma_{i,j}^t = \gamma_{i,j}$ for any $t = 2, ..., T$ and for any $j, i \in \{0, 1\}$, with the objective of evaluating the additional value of a non homogenous hidden Markov chain.

4.5 Parameters and hidden states estimation

The model parameters $\boldsymbol{\Psi} = (\mu, \boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$and the hidden states $\boldsymbol{s}^T$ for the non-homogenous models 1 and 2, and $\boldsymbol{\Psi} = (\mu, \boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ for the homogenous model 0 are numerically estimated using a bayesian approach via Markov chain Monte Carlo (MCMC) methods. For the homogenous model 0, all the parameters are sampled using the Gibbs algorithm, for the non-homogenous models 1 and 2 exception is made for the parameters $\boldsymbol{\alpha}$ of the transition probabilities that are sampled using a Metropolis-Hastings algorithm. The state sequence $\boldsymbol{s}^T$ is sampled using the *ff-bs* algorithm: forward filtering - backward sampling algorithm [95, 99].

4.5.1 Parameters prior distribution

The initial distribution of the hidden Markov chain is fixed as an uniform discrete distribution ($\delta_0 = \delta_1 = \frac{1}{2}$). The parameter independent prior distributions are set as:

- for all the models:
  - $\mu \sim N(\mu_M, \sigma_M^2)$ where $\mu_M = 50$ and $\sigma_M^2 = 10$ given that, empirically the rate of 50 ILI cases per 100,000 was the value above which Public Health officials in Portugal considered the start of the epidemic when baseline approaches were absent;
  - $\tau_i \sim Gamma(\alpha_\tau; \beta_\tau)$, where $\alpha_\tau = \beta_\tau = 0.5$, under the increasing order constraint ($\tau_0 > \tau_1$), for $i = 0, 1$;
  - $\theta_0 \sim N(\mu_\theta; \sigma_\theta^2)$ where $\mu_\theta = 0$ and $\sigma_\theta^2 = 10$;
  - $\boldsymbol{\theta}_1 \sim N_2(\boldsymbol{\mu}_\theta; \Sigma_\theta)$ where $\boldsymbol{\mu}_\theta = (0, 0)$ and $\Sigma_\theta = 10I_2$;
  - $\boldsymbol{\beta} = (\beta_1, \beta_2) \sim N_2(\boldsymbol{\mu}_B; \Sigma_B)$ where $\boldsymbol{\mu}_B = (0, 0)$ and $\Sigma_B = 5I_2$; where I_n is the identity matrix of dimension n.
- for Model 0: $\boldsymbol{\gamma}_0 = (\gamma_{0,0}, \gamma_{0,1})$ and $\boldsymbol{\gamma}_1 = (\gamma_{1,0}, \gamma_{1,1}) \sim Diriclet(\lambda_1, \lambda_2)$ where $\lambda_1 = \lambda_2 = 1$;
- for Model 1: $\boldsymbol{\alpha}_{0,1} = (\alpha_{0,1,0}, \alpha_{0,1,1})$ and $\boldsymbol{\alpha}_{1,0} = (\alpha_{1,0,0}, \alpha_{1,0,1}) \sim N_2(\mu_A; \Sigma_A)$, where $\mu_A = (0, 0)$ and $\Sigma_A = 10I_2$;
- for Model 2: $\boldsymbol{\alpha}_{0,1} = (\alpha_{0,1,0}, \alpha_{0,1,1}, \alpha_{0,1,2})$ and $\boldsymbol{\alpha}_{1,0} = (\alpha_{1,0,0}, \alpha_{1,0,1}, \alpha_{1,0,2}) \sim N_3(\mu_A; \Sigma_A)$, where $\mu_A = (0, 0, 0)$ and $\Sigma_A = 10I_3$.

4.5.2 Parameters posterior distribution

For the non homogenous models the posterior distribution of $\boldsymbol{\Psi}$ is then given by:

$$\begin{aligned} &\pi\left(\boldsymbol{\Psi}, \boldsymbol{s}^T | \boldsymbol{y}_{(+1)}^T, \boldsymbol{y}_0^T, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{\delta}\right) \\ &= \pi\left(\mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{s}^T | \boldsymbol{y}_{(+1)}^T, \boldsymbol{y}_{(0)0}^T, \boldsymbol{y}_{(0)1}^T, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{\delta}\right) \end{aligned} \tag{4.4}$$

$$\propto f\left(\boldsymbol{y}_{(+1)}^T | \mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{s}^T, \boldsymbol{y}_{(0)0}^T, \boldsymbol{y}_{(0)1}^T, \boldsymbol{C}\right) f\left(\boldsymbol{s}^T | \boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\delta}\right) p(\mu) p(\boldsymbol{\tau}) p(\theta_0) p(\boldsymbol{\theta}_1) p(\boldsymbol{\beta}) p(\boldsymbol{\alpha})$$

where $\boldsymbol{y}_{(+1)}^T = (y_{1(2)}, ..., y_{T(T+1)})'$ is the vector of the incidence rates estimated by Wednesday of week $t+1$, $\boldsymbol{s}^T = (s_1, ..., s_T)$ is the states vector of the hidden Markov

chain, $\boldsymbol{y}_{(0)0}^T = (y_{1(1)0}, ..., y_{T(T)0})'$ and $\boldsymbol{y}_{(0)1}^T = (y_{1(1)1}, ..., y_{T(T)1})'$ are the state specific vectors of the early estimate of the incidence rate calculated by Friday of week t, $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2)$ is a $2\times T$ matrix of the periodic component $\boldsymbol{c}_1 = (\cos(\frac{2\pi}{52}), ..., \cos(\frac{2t\pi}{52}), ..., \cos(\frac{2T\pi}{52}))'$ and $\boldsymbol{c}_2 = (\sin(\frac{2\pi}{52}), ..., \sin(\frac{2t\pi}{52}), ..., \sin(\frac{2T\pi}{52}))'$ and $\boldsymbol{Z} = (\boldsymbol{1}, \boldsymbol{y}_{(0)}^T)$ for Model 1 and $\boldsymbol{Z} = (\boldsymbol{1}, \boldsymbol{y}_{(0)}^T, \boldsymbol{v}_{(+1)}^{T-1})$ for Model 2 are respectively the vector and matrix of the covariates included in state transition probabilities matrix, where $\boldsymbol{v}_{(+1)}^{T-1} = (v_{0(1)}, ..., v_{T-1(T)})'$.

The likelihood can be factorized as:

$$f\left(\boldsymbol{y}_{(+1)}^T | \mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{s}^T, \boldsymbol{y}_{(0)0}^T, \boldsymbol{y}_{(0)1}^T, \boldsymbol{C}\right) =$$

$$= \prod_{t=1}^{T} f(y_{t(t+1)} | \mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, s_t, y_{t(t)0}, y_{t(t)1}, c_{1,t}, c_{2,t})$$

Here $f(y_{t(t+1)} | \mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, s_t, y_{t(t)0}, y_{t(t)1}, c_{1,t}, c_{2,t})$ is given by:

$$\sqrt{\frac{\tau_0}{2\pi}} \exp\left\{-\frac{\tau_0}{2}\left(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_0 y_{t(t)0}\right)^2\right\} \qquad \text{when} \qquad s_t = 0 \qquad \text{and}$$

$$\sqrt{\frac{\tau_1}{2\pi}} \exp\left\{-\frac{\tau_1}{2}\left(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_{1,1} y_{t(t)1} - \theta_{1,2} y_{t(t)1}^2\right)^2\right\} \qquad \text{if} \qquad s_t = 1.$$

The joint distribution of the hidden states is given by:

$$f(\boldsymbol{s}^T | \boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\delta}) = \delta_{s_1} \prod_{t=2}^{T} \gamma_{s_{t-1}, s_t}^t$$

For the homogenous HMM (Model 0) the state transition probabilities $\gamma_{i,j}^t = \gamma_{i,j}$, i.e. are obviously not indexed in time.

Finally, the posterior probability that a week t belongs to the epidemic state is given by the posterior mean of each s_t, being estimated as:

$$\hat{P}[S_t = 1] = \sum_{k=1}^{K} \frac{s_t^{(k)}}{K}$$

where $s_t^{(k)} \in \{0, 1\}$ are the sampled states in each iteration and K is the total number of iterations of the MCMC algorithm, after burn-in period and thinning. The classification of each week in the epidemic on non epidemic state is named as state decoding.

4.5.3 MCMC algorithm used for bayesian inference

The bayesian inference of the proposed HMM is done by sampling from the posterior distribution (equation 4.5).

More specifically, given the parameters vector:

$$\boldsymbol{\Psi}^{(k-1)} = \left(\mu^{(k-1)}, \boldsymbol{\tau}^{(k-1)}, \boldsymbol{\beta}^{(k-1)}, \theta_0^{(k-1)}, \boldsymbol{\theta}_1^{(k-1)}, \boldsymbol{\alpha}^{(k-1)}\right)$$

(for the homogenous model $\boldsymbol{\alpha}^{(k-1)}$ is substituted by $\boldsymbol{\gamma}^{(k-1)}$) and the hidden states $\boldsymbol{s}^{T(k-1)}$, generated in the k-1 iteration under the constraint $\tau_0^{(k-1)} > \tau_1^{(k-1)}$, the generic steps to obtain their values in the k iteration are described bellow, followed by the calculation of all the posterior distributions involved:

1. The state sequence $\boldsymbol{s}^{T(k)}$ is generated by the *ff-bs* algorithm;
2. The precisions $\tau_i, i = 0, 1$ are generated independently from their full conditional distribution which are Gamma distributions. Here the increasing order constrain must be applied if $\tau_0^{(k)} \leq \tau_1^{(k)}$. In this situation a permutation must applied to the state specific parameter values in order to change the values attributed to the non-epidemic state to the epidemic state and vice versa;
3. The full conditional distribution from which the signal value $\mu^{(k)}$ of the common part is generated from is a Normal distribution;
4. The vector of parameters $\boldsymbol{\beta}^{(k)}$ $\left(\beta_1^{(k)}, \beta_2^{(k)}\right)$ are generated from their full conditional posterior distribution which is a Multivariate Normal distribution;
5. The parameters $\theta_0^{(k)}$ and $\boldsymbol{\theta}_1^{(k)}$ are independently generated from their full conditional posterior distributions, which are Normal and Multinormal;
6. For the **homogenous model**: the i-th row of Γ^k matrix is generated from Dirichlet distribution. For the **non-homogenous models** the transition probabilities matrix parameters $\boldsymbol{\alpha_{0,1}}$ and $\boldsymbol{\alpha_{1,0}}$ are sampled using a random-walk Metropolis-Hastings algorithm.

The *forward filtering-backward sampling* algorithm

The *ff-bs* algorithm was developed in order to sample the sequence of hidden states $\boldsymbol{s}^T$ from the full conditional joint distribution of the states. This method was first proposed by Chib S in 1996 [99]. Here the rational of the algorithm is described.

Consider the full conditional joint distribution of the states:

$$P\left(\boldsymbol{s}^T|\boldsymbol{y}_{(+1)}^T,\boldsymbol{\Psi},\boldsymbol{D}\right)=P(s_T|\boldsymbol{y}_{(+1)}^T,\boldsymbol{\Psi},\boldsymbol{D})\prod_{t=1}^{T-1}P(s_t|\boldsymbol{y}_{(+1)}^T,\boldsymbol{s}_{t+1}^T\boldsymbol{\Psi},\boldsymbol{D})$$

where $\boldsymbol{D}=(\boldsymbol{y}_{(0)}^T,\boldsymbol{C},\boldsymbol{Z})$, $\boldsymbol{s}_{t+1}^T=(s_{t+1},...,s_T)$.

Given this expression the objective of the *ff-bs* algorithm is to sample each of the states s_t from $P(s_t|\boldsymbol{y}_{(+1)}^T,\boldsymbol{s}_{t+1}^T,\boldsymbol{\Psi},D)$ for $t=1,...,T-1$ and from $P(s_T|\boldsymbol{y}_{(+1)}^T,\boldsymbol{\Psi},\boldsymbol{D})$ for $t=T$.

Developing the generic term one gets,

$$P(s_t|\boldsymbol{y}_{(+1)}^T,\boldsymbol{s}_{t+1}^T,\boldsymbol{\Psi},\boldsymbol{D})=\frac{P(s_t,\boldsymbol{s}_{t+1}^T,\boldsymbol{y}_{(+1)}^t,\boldsymbol{y}_{t+1(+1)}^T,\boldsymbol{\Psi},\boldsymbol{D})}{P(\boldsymbol{s}_{t+1}^T,\boldsymbol{y}_{(+1)}^T,\boldsymbol{\Psi},\boldsymbol{D})}$$

$$\propto f(\boldsymbol{y}_{t+1(+1)}^T,\boldsymbol{s}_{t+1}^T|s_t,\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})P(s_t,\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})$$

$$=f(\boldsymbol{y}_{t+1(+1)}^T,\boldsymbol{s}_{t+1}^T|s_t,\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})P(s_t|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})f(\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})$$

$$\propto f(\boldsymbol{y}_{t+1(+1)}^T,\boldsymbol{s}_{t+1}^T|s_t,\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})P(s_t|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})$$

$$=f(\boldsymbol{y}_{t+1(+1)}^T,\boldsymbol{s}_{t+2}^T|s_{t+1},\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})P(s_{t+1}|s_t,\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})P(s_t|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})$$

$$\propto P(s_{t+1}|s_t,\boldsymbol{\Psi},\boldsymbol{D})P(s_t|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D}),$$

given that $f(\boldsymbol{y}_{t+1(+1)}^T,\boldsymbol{s}_{t+2}^T|s_t,s_{t+1},\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})$ does not depend on s_t.

So the generic term of the full conditional joint distribution of states can be derived from the product between $P(s_{t+1}|s_t,\boldsymbol{\Psi},\boldsymbol{D})$, the state transition probability, and $P(s_t|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})$.

In order to have a mass function this expression must be divided by a normalizing constant, so that:

$$P(s_t=i|\boldsymbol{y}_{(+1)}^T,s_{t+1}=j,\boldsymbol{\Psi},\boldsymbol{D})=\frac{P(s_t=i|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})P(s_{t+1}=j|s_t=i,\boldsymbol{\Psi},\boldsymbol{D})}{\sum_{k=0}^{1}P(s_t=k|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})P(s_{t+1}=j|s_t=k,\boldsymbol{\Psi},\boldsymbol{D})}$$

Where $i,j\in\{0,1\}$ and j is the state observed in the moment $t+1$.
From here it can be easily seen that sampling each state depends on the calculation of $P(s_t|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi},\boldsymbol{D})$.

In order to calculate this probability Chib S [99] proposed in 1996 a method based on a *update* step followed by a *prediction* one.

This method is supported by the following result,

$$P(s_t|\boldsymbol{y}_{(+1)}^{t},\boldsymbol{\Psi},\boldsymbol{D}) = P(s_t|y_{t(t+1)},\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D}) = \frac{P(s_t,y_{t(t+1)}|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D})}{f(y_{t(t+1)}|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D})}$$

,

$$\propto P(s_t,y_{t(t+1)}|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D}) = f(y_{t(t+1)}|\boldsymbol{y}_{(+1)}^{t-1},s_t,\boldsymbol{\Psi},\boldsymbol{D})P(s_t|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D}),$$

with mass function,

$$P(s_t=i|\boldsymbol{y}_{(+1)}^{t},\boldsymbol{\Psi},\boldsymbol{D}) = \frac{f(y_{t(t+1)}|\boldsymbol{y}_{(+1)}^{t-1},s_t=i,\boldsymbol{\Psi},\boldsymbol{D})P(s_t=i|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D})}{\sum_{k=0}^{1} f(y_{t(t+1)}|\boldsymbol{y}_{(+1)}^{t-1},s_t=k,\boldsymbol{\Psi},\boldsymbol{D})P(s_t=k|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D})}.$$

On the other hand from the law of total probability, one can assume that,

$$P(s_t|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D}) = \sum_{k=0}^{1} P(s_t|s_{t-1}=k,\boldsymbol{\Psi},\boldsymbol{D})P(s_{t-1}=k|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D}).$$

So, if the procedure is initialized considering that $P(s_1|\boldsymbol{\Psi},\boldsymbol{D})$ is given by $\boldsymbol{\delta}$, the probability distribution of the initial state, then, the probabilities $P(s_t|\boldsymbol{y}_{(+1)}^{t},\boldsymbol{\Psi},\boldsymbol{D})$ can be obtained from $t=2,...,T$, assuming that $P(s_{t-1}|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D})$ is available, as follows,

Prediction step:

$$P(s_t=i|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D}) = \sum_{k=0}^{1} P(s_t=i|s_{t-1}=k,\boldsymbol{\Psi},\boldsymbol{D})P(s_{t-1}=k|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D})$$

for $i\in\{0,1\}$,

Update step:

$$P(s_t=i|\boldsymbol{y}_{(+1)}^{t},\boldsymbol{\Psi},\boldsymbol{D}) = \frac{f(y_{t(t+1)}|\boldsymbol{y}^{t-1},\boldsymbol{\Psi},\boldsymbol{D},s_t=i)P(s_t=i|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D})}{\sum_{k=0}^{1} f(y_{t(t+1)}|\boldsymbol{y}_{(+1)}^{t-1},s_t=k,\boldsymbol{\Psi},\boldsymbol{D})P(s_t=k|\boldsymbol{y}_{(+1)}^{t-1},\boldsymbol{\Psi},\boldsymbol{D})}$$

for $i\in\{0,1\}$.

At $t=T$ the *forward filtering* phase of the algorithm ends and the *backward sampling* phase is initiated by sampling s_T from a Bernoulli distribution with probabilities $P(s_T=i|\boldsymbol{y}_{(+1)}^{T},\Psi,D)$ for $i\in\{0,1\}$.

This *backward sampling* phase corresponds to the procedure of sampling the hidden states vector $\boldsymbol{s}^T$ from their full conditional joint distribution. So, each of the

remaining $T-1$ states, is sampled going backwards (for $t = T-1, ..., 1$), from a Bernoulli distribution with probabilities:

$$P(s_t = i|\boldsymbol{y}_{(+1)}^T, s_{t+1} = j, \boldsymbol{\Psi}, \boldsymbol{D}) = \frac{P(s_t = i|\boldsymbol{y}_{(+1)}^t, \boldsymbol{\Psi}, \boldsymbol{D})P(s_{t+1} = j|s_t = i, \boldsymbol{\Psi}, \boldsymbol{D})}{\sum_{k=0}^{1} P(s_t = k|\boldsymbol{y}_{(+1)}^t, \boldsymbol{\Psi}, \boldsymbol{D})P(s_{t+1} = j|s_t = k, \boldsymbol{\Psi}, \boldsymbol{D})}$$

for $i \in \{0, 1\}$.

Posterior distribution of $\boldsymbol{\tau} = (\tau_0, \tau_1)$

The full conditional posteriori distribution of $\boldsymbol{\tau}$ is given by:

$$f(\boldsymbol{\tau}|\boldsymbol{y}_{(+1)}^T, \boldsymbol{s}^T, \boldsymbol{\Psi}, \boldsymbol{D}) =$$

$$= f(\tau_0|\boldsymbol{y}_{(+1)}^T, \boldsymbol{s}^T = 1, \boldsymbol{\Psi}, \boldsymbol{D}) f(\tau_1|\boldsymbol{y}_{(+1)}^T, \boldsymbol{s}^T = 0, \boldsymbol{\Psi}, \boldsymbol{D})$$

where $\boldsymbol{s}^T = i$ for $i \in \{0, 1\}$, represents that each distributions is conditioned on the subsets of $\boldsymbol{y}_{(+1)}^T$ and $\boldsymbol{D}$ where $\{t \geq 1 : s_t = i\}$.

Lets now consider that,

$$f(\tau_i|\boldsymbol{y}_{(+1)}^T, \boldsymbol{s}^T = i, \boldsymbol{\Psi}, \boldsymbol{D}) \propto f(\boldsymbol{y}_{(+1)}^T|\boldsymbol{s}^T = i, \boldsymbol{\Psi}, \boldsymbol{D}) p(\tau_i)$$

were $p(\tau_i)$ is the priori distribution of the parameters, τ_i defined as $Gamma(\alpha_\tau, \beta_\tau)$.

So for $i = 0$, considering that $y_{t(t+1)}|s_t = 0, \boldsymbol{\Psi}, \boldsymbol{D} \sim N(\mu + \beta_1 c_{1,t} + \beta_2 c_{2,t} + \theta_0 y_{t(t)0}, \tau_0)$ for $\{t \geq 1 : s_t = 0\}$ the posterior distribution of τ_0 will be Gamma with parameters,

$$\alpha_0 = \alpha_\tau + \frac{T_0}{2}$$

$$\beta_0 = \beta_\tau + \frac{1}{2} \sum_{\{t \geq 1 : s_t = 0\}} \left(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_0 y_{t(t)0}\right)^2,$$

where T_0 is the number of observations that corresponds to the state 0.

This results comes from the fact that the Gamma distribution is a conjugate prior for the Normal likelihood and $\boldsymbol{\Psi}$ is known.

Identically applying the same rule for $i = 1$, shows that the posterior distribution of τ_1 is Gamma distributed with parameters:

$$\alpha_1 = \alpha_\tau + \frac{T_1}{2}.$$

$$\beta_1 = \beta_\tau + \frac{1}{2} \sum_{\{t \geq 1 : s_t = 1\}} \left(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_{1,1} y_{t(t)1} - \theta_{1,1} y_{t(t)1}^2\right)^2,$$

where T_1 is the number observations that corresponds to the state 1.

Posterior distribution of μ

The full conditional posteriori distribution of μ is given by:

$$f(\mu|\boldsymbol{y}_{(+1)}^T, \boldsymbol{s}^T, \boldsymbol{\Psi}, \boldsymbol{D})$$

$$\propto f(\boldsymbol{y}_{(+1)}^T|\boldsymbol{s}^T, \boldsymbol{\Psi}, \boldsymbol{D})p(\mu)$$

were $p(\mu)$ is the priori distribution of μ defined as Normal with mean μ_M and variance σ_M^2.

Developing this expression:

$$f(\boldsymbol{y}_{(+1)}^T|\boldsymbol{s}^T, \boldsymbol{\Psi}, \boldsymbol{D})p(\mu) =$$
$$= \prod_{\{t\geq 1:s_t=0\}} f(y_{t(t+1)}|s_t = 0, \boldsymbol{\Psi}, \boldsymbol{D}) \prod_{\{t\geq 1:s_t=1\}} f(y_{t(t+1)}|s_t = 1, \boldsymbol{\Psi}, \boldsymbol{D})p(\mu) \propto$$
$$\propto \prod_{\{t\geq 1:s_t=0\}} \exp\{-\tfrac{\tau_0}{2}(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_0 y_{t(t)0})^2\}$$
$$\prod_{\{t\geq 1:s_t=1\}} \exp\{-\tfrac{\tau_0}{2}(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_{1,1} y_{t(t)1} - \theta_{1,2} y_{t(t)1}^2)^2\}$$
$$\exp\{-\tfrac{\tau_M}{2}(\mu - \mu_M)^2\} \propto$$
$$\propto \prod_{\{t\geq 1:s_t=0\}} \exp\{-\tfrac{\tau_0}{2}(\mu^2 - 2\mu\tilde{y}_{t(t+1)}^0)\} \prod_{\{t\geq 1:s_t=1\}} \exp\{-\tfrac{\tau_1}{2}(\mu^2 - 2\mu\tilde{y}_{t(t+1)}^1)\}$$
$$\exp\{-\tfrac{\tau_M}{2}(\mu^2 - 2\mu\mu_M)\} = (*)$$

where

$$\tilde{y}_{t(t+1)}^0 = y_{t(t+1)} - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_0 y_{t(t)0}$$

for $\{t \geq 1 : s_t = 0\}$ and

$$\tilde{y}_{t(t+1)}^1 = y_{t(t+1)} - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_{1,1} y_{t(t)1} - \theta_{1,2} y_{t(t)1}^2$$

for $\{t \geq 1 : s_t = 1\}$.
Continuing the development:

$$(*) = \exp\{-\tfrac{\tau_0}{2}(T_0\mu^2 - 2\mu \textstyle\sum_{\{t\geq 1:s_t=0\}} \tilde{y}_{t(t+1)}^0) \; - \tfrac{\tau_1}{2}(T_1\mu^2 - 2\mu \textstyle\sum_{\{t\geq 1:s_t=1\}} \tilde{y}_{t(t+1)}^1)\}$$
$$\exp\{-\tfrac{\tau_M}{2}(\mu^2 - 2\mu\mu_M)\} =$$

where T_0 and T_1 is the number of observations that corresponds respectively to the

state 0 (non epidemic) and 1 (epidemic).

$$= \exp\{-\tfrac{\mu^2}{2}(\tau_0 T_0 + \tau_1 T_1) + \mu\tau_0 \textstyle\sum_{\{t\geq 1:s_t=0\}} \tilde{y}^0_{t(t+1)} + \mu\tau_1 \sum_{\{t\geq 1:s_t=1\}} \tilde{y}^1_{t(t+1)}\}$$

$$\exp\{-\tfrac{\tau_M}{2}(\mu^2 - 2\mu\mu_M)\} =$$

$$\exp\{-\tfrac{\mu^2}{2}(\tau_0 T_0 + \tau_1 T_1 + \tau_M) + \mu(\tau_0 \textstyle\sum_{\{t\geq 1:s_t=0\}} \tilde{y}^0_{t(t+1)} + \tau_1 \sum_{\{t\geq 1:s_t=1\}} \tilde{y}^1_{t(t+1)}) + \mu_M\tau_M\} =$$

$$= \exp\{-\tfrac{\tau_0 T_0+\tau_1 T_1+\tau_M}{2}\left[\mu^2 - 2\mu\left(\tfrac{\tau_0 \sum_{\{t\geq 1:s_t=0\}} \tilde{y}^0_{t(t+1)}+\tau_1 \sum_{\{t\geq 1:s_t=1\}} \tilde{y}^1_{t(t+1)}+\mu_M\tau_M}{\tau_0 T_0+\tau_1 T_1+\tau_M}\right)\right]\} \propto$$

$$\propto \exp\{-\tfrac{\tau_0 T_0+\tau_1 T_1+\tau_M}{2}\left(\mu - \tfrac{\tau_0 \sum_{\{t\geq 1:s_t=0\}} \tilde{y}^0_{t(t+1)}+\tau_1 \sum_{\{t\geq 1:s_t=1\}} \tilde{y}^1_{t(t+1)}+\mu_M\tau_M}{\tau_0 T_0+\tau_1 T_1+\tau_M}\right)^2\}$$

From which we can conclude that the full conditional posterior distribution of μ is Normal distributed with mean:

$$\frac{\tau_0 \sum_{\{t\geq 1:s_t=0\}} \tilde{y}^0_{t(t+1)} + \tau_1 \sum_{\{t\geq 1:s_t=1\}} \tilde{y}^1_{t(t+1)} + \mu_M\tau_M}{\tau_0 T_0 + \tau_1 T_1 + \tau_M}$$

and precision $\tau_0 T_0 + \tau_1 T_1 + \tau_M$.

Posterior distribution of $\boldsymbol{\beta} = (\beta_1, \beta_2)$

The full conditional posteriori distribution of $\boldsymbol{\beta}$ is given by:

$$f(\boldsymbol{\beta}|\boldsymbol{y}^T_{(+1)}, \boldsymbol{s}^T, \boldsymbol{\Psi}, \boldsymbol{D})$$

$$\propto f(\boldsymbol{y}^T_{(+1)}|\boldsymbol{s}^T, \boldsymbol{\Psi}, \boldsymbol{D}) p(\boldsymbol{\beta})$$

where $p(\boldsymbol{\beta})$ is the priori distribution of $\boldsymbol{\beta}$ defined as Multivariate Normal with mean $\boldsymbol{\mu}_B$ and covariance matrix Σ_B. Note that,

$$f(\boldsymbol{y}^T_{(+1)}|\boldsymbol{s}^T, \boldsymbol{\Psi}, \boldsymbol{D}) p(\boldsymbol{\beta}) \propto$$

$$\propto \textstyle\prod_{\{t\geq 1:s_t=0\}} \exp\{-\frac{\tau_0}{2}(y_{t(t+1)} - \mu - \theta_0 y_{t(t)0} - \beta_1 c_{1,t} - \beta_2 c_{2,t})^2\}$$

$$\textstyle\prod_{\{t\geq 1:s_t=1\}} \exp\{-\frac{\tau_1}{2}(y_{t(t+1)} - \mu - \theta_{1,1} y_{t(t)1} - \theta_{1,2} y^2_{t(t)1} - \beta_1 c_{1,t} - \beta_2 c_{2,t})^2\}$$

$$\exp\left\{-\tfrac{1}{2}(\boldsymbol{\beta} - \mu_B)'\Sigma_B^{-1}(\boldsymbol{\beta} - \mu_B)\right\}$$

can be rewriten in the following matrix form:

$$= \exp\left\{-\frac{1}{2}(\tilde{y}^T_{(+1)} - \boldsymbol{C\beta})'\boldsymbol{Q}(\tilde{y}^T_{(+1)} - \boldsymbol{C\beta})\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mu_B)'\Sigma_B^{-1}(\boldsymbol{\beta} - \mu_B)\right\} \propto$$

where $\boldsymbol{Q}$ is the diagonal matrix whose the tth term is τ_0 if $s_t = 0$ or τ_1 if $s_t = 1$. On the other hand $\tilde{y}^T_{(+1)}$ is the vector with generic terms $\left(y_{t(t+1)} - \mu - \theta_0 y_{t(t)0}\right)$ if $s_t = 0$ or $\left(y_{t(t+1)} - \mu - \theta_{1,1} y_{t(t)1} - \theta_{1,2} y^2_{t(t)1}\right)$ if $s_t = 1$.
Continuing the development produces,

$$\propto \exp\left\{-\tfrac{1}{2}(\tilde{y}^{T'}_{(+1)}\boldsymbol{Q}\tilde{y}^T_{(+1)} - 2\boldsymbol{\beta}'\boldsymbol{C}'\boldsymbol{Q}\tilde{y}^T_{(+1)} + \boldsymbol{\beta}'\boldsymbol{C}'\boldsymbol{Q}\boldsymbol{C}\boldsymbol{\beta}) - \tfrac{1}{2}(\boldsymbol{\beta}'\Sigma_B^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\Sigma_B^{-1}\mu_B + \mu_B'\Sigma_B^{-1}\mu_B)\right\} \propto$$
$$\propto \exp\left\{-\tfrac{1}{2}\left[\boldsymbol{\beta}'\left(\boldsymbol{C}'\boldsymbol{Q}\boldsymbol{C} + \Sigma_B^{-1}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}'\left(\boldsymbol{C}'\boldsymbol{Q}\tilde{y}^T_{(+1)} + \Sigma_B^{-1}\mu_B\right)\right]\right\}$$

So considering that $\Lambda_{\boldsymbol{\beta}} = \boldsymbol{C}'\boldsymbol{Q}\boldsymbol{C} + \Sigma_B^{-1}$ and $\mu_{\boldsymbol{\beta}} = \Lambda_{\boldsymbol{\beta}}^{-1}(\boldsymbol{C}'\boldsymbol{Q}\tilde{y}^T_{(+1)} + \Sigma_B^{-1}\mu_B)$ the full conditional posterior distribution of $\boldsymbol{\beta}$ will be proportional to:

$$\propto \exp\left\{(\boldsymbol{\beta} - \mu_{\boldsymbol{\beta}})'\Lambda_{\boldsymbol{\beta}}(\boldsymbol{\beta} - \mu_{\boldsymbol{\beta}})\right\}$$

This means that the full conditional posterior distribution of $\boldsymbol{\beta}$ is Multivariate Normal distributed with mean:

$$\mu_{\boldsymbol{\beta}} = (\Sigma_B^{-1} + \boldsymbol{C}'\boldsymbol{Q}\boldsymbol{C})^{-1}(\Sigma_B^{-1}\mu_B + \boldsymbol{C}'\boldsymbol{Q}\tilde{y}^T_{(+1)})$$

and matrix precision:

$$\Lambda_{\boldsymbol{\beta}} = (\Sigma_B^{-1} + \boldsymbol{C}'\boldsymbol{Q}\boldsymbol{C})$$

Posterior distribution of θ_0 and $\boldsymbol{\theta}_1 = (\theta_{1,1}, \theta_{1,2})$

The full conditional posteriori distribution of θ_0 is given by:

$$f(\theta_0|\boldsymbol{y}^T_{(+1)}, \boldsymbol{s}^T = 0, \boldsymbol{\Psi}, \boldsymbol{D})$$
$$\propto f(\boldsymbol{y}^T_{(+1)}|\boldsymbol{s}^T = 1, \boldsymbol{\Psi}, \boldsymbol{D})p(\theta_0),$$

were $p(\theta_0)$ is the prior distribution of θ_0, defined as Normal with mean μ_θ and precision $\tau_\theta = \sigma_\theta^{-2}$,

$$\propto \prod_{\{t\geq 1: s_t = 0\}} \exp\left\{-\tfrac{\tau_0}{2}(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_0 y_{t(t)0})^2\right\} \exp\left\{-\tfrac{\tau_\theta}{2}(\theta_0 - \mu_\theta)^2\right\}$$
$$\propto \exp\left\{-\tfrac{\tau_0}{2}\sum_{\{t\geq 1: s_t = 1\}}(\theta_0 y_{t(t)0})^2 - 2(\theta_0 y_{t(t)0}(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t}))\right\} \exp\left\{-\tfrac{\tau_\theta}{2}(\theta_0^2 - 2\mu_\theta\theta_0)\right\}$$
$$= \exp\left\{-\frac{\tau_0\sum_{\{t\geq 1: s_t = 0\}} y^2_{t(t)0} - \tau_\theta}{2}\left[\theta_0^2 - \frac{2\theta_0\left(\tau_0\sum_{\{t\geq 1: s_t = 0\}} y_{t(t)0}(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t}) + \mu_\theta\tau_\theta\right)}{\tau_0\sum_{\{t\geq 1: s_t = 0\}} y^2_{t(t)0} - \tau_\theta}\right]\right\}$$
$$\propto \exp\left\{-\frac{\tau_0\sum_{\{t\geq 1: s_t = 0\}} y^2_{t(t)0} - \tau_\theta}{2}\left(\theta_0 - \frac{\tau_0\sum_{\{t\geq 1: s_t = 0\}} y_{t(t)0}(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t}) + \mu_\theta\tau_\theta}{\tau_0\sum_{\{t\geq 1: s_t = 0\}} y^2_{t(t)0} - \tau_\theta}\right)^2\right\}$$

Concluding that the full conditional posterior distribution of θ_0 is Normal distributed with mean,

$$\frac{\tau_0 \sum_{\tau_0\{t\geq 1:s_t=0\}} y_{t(t)0}(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t}) + \mu_\theta \tau_\theta}{\tau_0 \sum_{\{t\geq 1:s_t=0\}} y_{t(t)0}^2 - \tau_\theta},$$

and precision,

$$\tau_0 \sum_{\{t\geq 1:s_t=0\}} y_{t(t)0}^2 - \tau_\theta.$$

The full conditional posteriori distribution of $\boldsymbol{\theta}_1 = (\theta_{1,1}, \theta_{1,2})$ is given by:

$$\begin{aligned} &f(\boldsymbol{\theta_1}|\boldsymbol{y}_{(+1)}^T, \boldsymbol{s}^T = 1, \boldsymbol{\Psi}, \boldsymbol{D}) \\ &\propto f(\boldsymbol{y}_{(+1)}^T|\boldsymbol{s}^T = 1, \boldsymbol{\Psi}, \boldsymbol{D}) p(\boldsymbol{\theta_1}) \end{aligned}$$

were $p(\boldsymbol{\theta}_1)$ is the prior distribution of $\boldsymbol{\theta}_1$, defined as Multivariate Normal with mean $\boldsymbol{\mu}_\theta$ and covariance matrix Σ_θ. Notice that,

$$\begin{aligned} &f(\boldsymbol{y}_{(+1)}^T|\boldsymbol{s}^T = 1, \boldsymbol{\Psi}, \boldsymbol{D}) p(\boldsymbol{\theta_1}) \propto \\ &\propto \prod_{\{t\geq 1:s_t=1\}} \exp\{-\tfrac{\tau_1}{2}(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_{1,1} y_{t(t)1} - \theta_{1,2} y_{t(t)1}^2)^2\} \\ &\exp\left\{-\tfrac{1}{2}(\boldsymbol{\theta_1} - \boldsymbol{\mu}_\theta)' \Sigma_\theta^{-1} (\boldsymbol{\theta_1} - \boldsymbol{\mu}_\theta)\right\} \end{aligned}$$

can be rewriten in the following matrix form,

$$\exp\left\{-\frac{1}{2}(\tilde{y}_{(+1)}^T - \tilde{\boldsymbol{Y}}_{(0)}\boldsymbol{\theta_1})' \boldsymbol{Q_1} (\tilde{y}_{(+1)}^T - \tilde{\boldsymbol{Y}}_{(0)}\boldsymbol{\theta_1})\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta_1} - \boldsymbol{\mu}_\theta)' \Sigma_\theta^{-1} (\boldsymbol{\theta_1} - \boldsymbol{\mu}_\theta)\right\}$$

where $\boldsymbol{Q_1}$ is the diagonal matrix which the tth term is τ_1 if $s_t = 1$ and zero otherwise. On the other hand, $\tilde{y}_{(+1)}^T$ is the vector with generic terms $\left(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t}\right)$ if $s_t = 1$ and zero otherwise, and $\tilde{\boldsymbol{Y}}_{(0)} = (y_{(0)}^T, (y_{(0)}^T)^2)$.

Following the same calculations applied to obtain the full conditional posterior distribution of $\boldsymbol{\beta}$, one can conclude that the full conditional posterior distribution of $\boldsymbol{\theta_1}$ is Multivariate Normal distributed with mean,

$$(\Sigma_\theta^{-1} + \tilde{\boldsymbol{Y}}'_{(0)} \boldsymbol{Q_1} \tilde{\boldsymbol{Y}}_{(0)})^{-1} (\Sigma_\theta^{-1} \boldsymbol{\mu}_\theta + \tilde{\boldsymbol{Y}}'_{(0)} \boldsymbol{Q_1} \tilde{y}_{(+1)}^T)$$

and matrix precision:

$$(\Sigma_\theta^{-1} + \tilde{\boldsymbol{Y}}'_{(0)} \boldsymbol{Q_1} \tilde{\boldsymbol{Y}}_{(0)}).$$

Posterior distribution of the parameters associated with the state transition probabilities matrix

Consider first the **homogenous** case. The full conditional posteriori distribution of $\mathbf{\Gamma}$ is given by:

$$f(\mathbf{\Gamma}|s^T, \boldsymbol{\delta}) =$$
$$= f(\boldsymbol{\gamma}_0|s^T = 0, \boldsymbol{\delta}) f(\boldsymbol{\gamma}_1|s^T = 1, \boldsymbol{\delta})$$

where $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are respectively the first and second row of the state transition probabilities matrix. So the posterior distribution of $\boldsymbol{\gamma}_0$ is given by:

$$f(\boldsymbol{\gamma}_0|s^T = 0, \delta) \propto \pi(s^T = 0|\boldsymbol{\gamma}_0, \delta) p(\boldsymbol{\gamma}_0)$$

where $p(\boldsymbol{\gamma}_0)$, the prior distribution of $\boldsymbol{\gamma_0}$, is defined as $Dirichlet(\lambda_1, \lambda_2)$. Then,

$$\propto \delta_{s_1} \prod_{\{t \geq 2 : s_{t-1} = 0\}} \gamma_{s_{t-1}, s_t} \frac{1}{B(\lambda_1, \lambda_2)} \gamma_{0,0}^{\lambda_1 - 1} \gamma_{0,1}^{\lambda_2 - 1}$$

where $B(\lambda_1, \lambda_2)$ is the beta function and

$$= \delta_{s_1} \frac{1}{B(\lambda_1, \lambda_2)} \gamma_{0,0}^{n_{0,0}} \gamma_{0,1}^{n_{0,1}} \gamma_{0,0}^{\lambda_1 - 1} \gamma_{0,1}^{\lambda_2 - 1}$$

where $n_{0,0}$ and $n_{0,1}$ are respectively the number of pairs of observations with states non epidemic followed by non epidemic, and states non epidemic followed by the epidemic. So,

$$f(\boldsymbol{\gamma}_0|s^T = 0, \delta) \propto \frac{1}{B(n_{0,0} + \lambda_1, n_{0,1} + \lambda_2)} \gamma_{0,0}^{n_{0,0} + \lambda_1 - 1} \gamma_{0,1}^{n_{0,1} + \lambda_1 - 1}$$

Showing that the full conditional posterior distribution of $\boldsymbol{\gamma}_0$ is $Dirichlet(n_{0,0} + \lambda_1, n_{0,1} + \lambda_2)$. For parameter $\boldsymbol{\gamma}_1$, the second row of the state transition probabilities matrix, we conclude analogously that the full conditional posterior distribution is $Dirichlet(n_{1,0} + \lambda_1, n_{1,1} + \lambda_2)$ where $n_{1,0}$ and $n_{1,1}$ are respectively the number of pairs of observations with states epidemic followed by non epidemic and states epidemic followed by epidemic.

Lets now focus on the **non-homogenous** case. For model 1 and 2 the parameters $\boldsymbol{\alpha}_{0,1}$ and $\boldsymbol{\alpha}_{1,0}$ are sampled using a random walk Metropolis-Hastings algorithm. So considering that $\boldsymbol{\alpha}_{i,j}^{(k-1)}$ are the values of the parameters sampled in the $k-1$ iteration of the algorithm, the parameters of the k iteration are independently sampled from a random walk:

$$\boldsymbol{\alpha}_{i,j}^{(k)} = \boldsymbol{\alpha}_{i,j}^{(k-1)} + U_A$$

with $i, j = 0, 1$ and $i \neq j$, and U_A is generated a from Multivariate Normal distribution with parameters $\boldsymbol{\mu}_A$ and covariance matrix Σ_A. So, in each iteration the vectors $\boldsymbol{\alpha}_{0,1}^{(k)}$ and $\boldsymbol{\alpha}_{1,0}^{(k)}$ are accepted respectively with probability:

$$A(\boldsymbol{\alpha}_{0,1}^{(k)}, \boldsymbol{\alpha}_{0,1}^{(k-1)}) = \min\left(1; \frac{\pi\left(\boldsymbol{\alpha}_{0,1}^{(k)} | \boldsymbol{s}^{T(k)}, \boldsymbol{Z}, \boldsymbol{\delta}\right)}{\pi\left(\boldsymbol{\alpha}_{0,1}^{(k-1)} | \boldsymbol{s}^{T(k)}, \boldsymbol{Z}, \boldsymbol{\delta}\right)}\right)$$

and

$$A(\boldsymbol{\alpha}_{1,0}^{(k)}, \boldsymbol{\alpha}_{1,0}^{(k-1)}) = \min\left(1; \frac{\pi\left(\boldsymbol{\alpha}_{1,0}^{(k)} | \boldsymbol{s}^{T(k)}, \boldsymbol{Z}, \boldsymbol{\delta}\right)}{\pi\left(\boldsymbol{\alpha}_{1,0}^{(k-1)} | \boldsymbol{s}^{T(k)}, \boldsymbol{Z}, \boldsymbol{\delta}\right)}\right).$$

Here

$$\pi\left(\boldsymbol{\alpha}_{0,1}^{(k)} | \boldsymbol{s}^{T(k)}, \boldsymbol{Z}, \boldsymbol{\delta}\right) \propto$$

$$\propto \delta_{s_1^{(k)}} \prod_{\{t \geq 2 : s_{t-1}^{(k)} = 0\}} \gamma_{s_{t-1}^{(k)}, s_t^{(k)}}^{t(k)} \exp\{-1/2(\boldsymbol{\alpha}_{0,1}^{(k)} - \mu_A)' \Sigma_A^{-1} (\boldsymbol{\alpha}_{0,1}^{(k)} - \mu_A)\}$$

and

$$\pi\left(\boldsymbol{\alpha}_{1,0}^{(k)} | \boldsymbol{s}^{T(k)}, \boldsymbol{Z}, \boldsymbol{\delta}\right) \propto$$

$$\propto \delta_{s_1^{(k)}} \prod_{\{t \geq 2 : s_{t-1}^{(k)} = 1\}} \gamma_{s_{t-1}^{(k)}, s_t^{(k)}}^{t(k)} \exp\{-1/2(\boldsymbol{\alpha}_{1,0}^{(k)} - \mu_A)' \Sigma_A^{-1} (\boldsymbol{\alpha}_{1,0}^{(k)} - \mu_A)\}$$

where

$$\gamma_{j,i}^{t(k)} = \frac{\exp(\boldsymbol{\alpha}_{j,i}^{(k)} \boldsymbol{Z}_t)}{(1 + \exp(\boldsymbol{\alpha}_{j,i}^{(k)} \boldsymbol{Z}_t))}$$

and $\boldsymbol{Z}_t = (1, y_{t(t)})$ for model 1 and $\boldsymbol{Z}_t = (1, y_{t(t)}, v_{t-1(t)})$ for model 2.

4.5.4 Nowcasting weekly influenza activity states and ILI rates

In order to simulate the online performance of the models, the influenza activity state and the correspondent ILI rate of each week, from week 40/2010 to week 16/2011, were nowcasted. Each proposed model was sequentially fitted to all the data known until the respective previous week to nowcast the following. For example, week 40/2010 was nowcasted from week 40/2008 to week 39/2010, week 41/2010 was nowcasted from week 40/2008 to week 40/2010, and so on.

More specifically, consider that we are on Friday of week T, knowing all ILI rates and ILI cases tested positive for influenza until week $T-1$, and we want to nowcast week T, using previous information and the incomplete estimate of the ILI rate of week T. The probability that week T belongs to the epidemic influenza activity state is estimated by:

$$\hat{P}[S_T = 1] = \sum_{k=1}^{K} \frac{\hat{P}[S_T = 1|\mathbf{\Psi}^{(k)}]}{K} \tag{4.5}$$

where $\hat{P}[S_T = 1|\mathbf{\Psi}^{(k)}] = \hat{P}[S_{T-1} = 1|\mathbf{\Psi}^{(k)}]\gamma_{1,1}^{T(k)} + \hat{P}[S_{T-1} = 0|\mathbf{\Psi}^{(k)}]\gamma_{0,1}^{T(k)}$, k is the iteration step of the MCMC algorithm, $\hat{P}[S_{T-1} = 1|\mathbf{\Psi}^{(k)}]$ is the probability that week $T-1$ belongs to the epidemic state sampled in the kth iteration by the *ff-bs* algorithm and $\gamma_{j,i}^{T(k)}$ with $j, i \in \{0, 1\}$ represent the transition probabilities sampled in the kth iteration of the MCMC algorithm. For model 0 the transition probabilities sampled in each iteration are not time-dependent $\gamma_{j,i}^{(k)}$, but for the non homogenous models the transition probabilities are estimated using the known covariates at moment T and the parameters $\boldsymbol{\alpha}_{j,i}^{(k)}$ sampled in iteration k,

$$\gamma_{j,i}^{T(k)} = \frac{\exp(\boldsymbol{\alpha}_{j,i}^{(k)}\boldsymbol{Z}_T)}{(1 + \exp(\boldsymbol{\alpha}_{j,i}^{(k)}\boldsymbol{Z}_T))}$$

where $\boldsymbol{Z}_T = (1, y_{T(T)})$ for model 1 and $\boldsymbol{Z}_T = (1, y_{T(T)}, v_{T-1(T)})$ for model 2.

The nowcasting of the ILI rate for week T to be reported on week $T+1$ is estimated by:

$$\hat{y}_{T(T+1)} = \sum_{k=1}^{K} \frac{y_{T(T+1)|s_T^{(k)}=1}^{(k)} \hat{P}[S_T = 1|\Psi^{(k)}] + y_{T(T+1)|s_T^{(k)}=0}^{(k)}(1 - \hat{P}[S_T = 1|\Psi^{(k)}])}{K} \tag{4.6}$$

where $y_{T(T+1)|s_T^{(k)}=1}^{(k)}$ is sampled from $f(y_{T(T+1)}|\mu^{(k)}, \boldsymbol{\beta}^{(k)}, \tau_0^{(k)}, \theta_0^{(k)}, y_{T(T)0}, c_{1,T}, c_{2,T})$ and $y_{T(T+1)|s_T^{(k)}=0}^{(k)}$ is sampled from $f(y_{T(T+1)}|\mu^{(k)}, \boldsymbol{\beta}^{(k)}, \tau_1^{(k)}, \boldsymbol{\theta}_1^{(k)}, y_{T(T)1}, c_{1,T}, c_{2,T})$.

4.5.5 Model comparison and Marginal likelihood estimation

The Bayes factor was used in order to compare the models fit to data. For this purpose marginal likelihoods were computed numerically using the method presented by Chib 1995 [104], for the homogenous model, given that all parameters are sampled using the Gibbs algorithm, and the method presented by Chib and Jeliazkov 2001

[105] for the non homogenous models, since the transition probabilities parameters are sampled via the Metropolis-Hastings algorithm.

In both approaches the natural logarithm of the marginal likelihood is numericaly estimated via a MCMC algorithm in a special point $\boldsymbol{\Psi}^*$ of the parameters space, the posterior mode of $\boldsymbol{\Psi}$.

Lets consider the estimate of the marginal likelihood:

$$\hat{f}(y_{(+1)}^T|\boldsymbol{D},\boldsymbol{\delta}) = \frac{f(y_{(+1)}^T|\boldsymbol{\Psi}^*,\boldsymbol{D},\boldsymbol{\delta})p(\boldsymbol{\Psi}^*)}{\hat{\pi}(\boldsymbol{\Psi}^*|y_{(+1)}^T,\boldsymbol{D},\boldsymbol{\delta})},$$

and its natural logarithm

$$\ln(\hat{f}(y_{(+1)}^T|\boldsymbol{D},\boldsymbol{\delta}) = \ln(2) + \ln f(y_{(+1)}^T|\boldsymbol{\Psi}^*,\boldsymbol{D},\boldsymbol{\delta}) + \ln p(\boldsymbol{\Psi}^*) - \ln \hat{\pi}(\boldsymbol{\Psi}^*|y_{(+1)}^T,\boldsymbol{D},\boldsymbol{\delta}) \quad (4.7)$$

where the component $\ln(2)$ is added by suggestion of Neil 1999 [106], and represents the constraint imposed to achieve the identifiability of the states(in our case $\tau_0 > \tau_1$). This component is $\ln(m!)$ in the general case of an m-states model.

In order to estimate this logarithm, the MCMC algorithm presented in section 4.5.3 is executed for K iterations and parameters runs used to calculate $\boldsymbol{\Psi}^* = (\boldsymbol{\alpha}_{0,1}^*, \boldsymbol{\alpha}_{1,0}^*, \mu^*, \boldsymbol{\tau}^*, \theta_0^*, \theta_1^*, \boldsymbol{\beta}^*)$. In the next steps the second and third therm of equation 4.7 are calculated.

The seconde term of 4.7 is estimated estimated by,

$$\ln f(y_{(+1)}^T|\boldsymbol{\Psi}^*,\boldsymbol{D},\boldsymbol{\delta}) = \sum_{t=1}^{T} \ln \left[\sum_{i=0}^{1} f(y_{(+1)}^T|\boldsymbol{\Psi}^*,\boldsymbol{D},\boldsymbol{\delta},s_t=i) P(s_t=i|\boldsymbol{y}_{(+1)}^t,\boldsymbol{\Psi}^*,\boldsymbol{D}) \right]$$

where $P(s_t = i|\boldsymbol{y}_{(+1)}^t, \boldsymbol{\Psi}^*, \boldsymbol{D})$ is the filtered probability obtained in the *ff-bs* algorithm.

The third term of 4.7 is estimated in the non homogenous model by:

$$\ln p(\boldsymbol{\Psi}^*) = \ln p(\mu^*) + \ln p(\boldsymbol{\beta}^*) + \ln p(\tau_0^*) + \ln p(\tau_1^*) + \ln p(\theta_0^*) + \ln p(\boldsymbol{\theta}_1^*) + \ln p(\boldsymbol{\alpha}_{0,1}^*) + \ln p(\boldsymbol{\alpha}_{1,0}^*).$$

For the homogenous model the last two terms are replaced by $\ln p(\boldsymbol{\gamma}_0)$ and $\ln p(\boldsymbol{\gamma}_1)$).

The fourth term of 4.7 is calculated as follows:

$$
\begin{aligned}
\ln \quad & \hat{\pi}(\boldsymbol{\Psi}^*|\boldsymbol{y}_{(+1)}^T, \boldsymbol{D}) \qquad (4.8)\\
= \quad & \ln\left[K^{-1}\sum_{k=1}^{K}\left(A(\boldsymbol{\alpha}_{0,1}^{(k)}, \boldsymbol{\alpha}_{0,1}^*)q(\boldsymbol{\alpha}_{0,1}^*, \boldsymbol{\alpha}_{0,1}^{(k)})A(\boldsymbol{\alpha}_{1,0}^{(k)}, \boldsymbol{\alpha}_{1,0}^*)q(\boldsymbol{\alpha}_{1,0}^*, \boldsymbol{\alpha}_{1,0}^{(k)})\right)\right]\\
- \quad & \ln\left[K^{-1}\sum_{k=1}^{K}A(\alpha_{0,1}^*, \alpha_{0,1}^{(k)})A(\alpha_{1,0}^*, \alpha_{1,0}^{(k)})\right]\\
+ \quad & \ln\left[K^{-1}\sum_{k=1}^{K}\pi(\mu^*|\boldsymbol{y}_{(+1)}^T, \boldsymbol{D}, \boldsymbol{\alpha}^*, \boldsymbol{\tau}^{(k)}, \boldsymbol{\theta}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{s}^{T(k)})\right]\\
+ \quad & \ln\left[K^{-1}\sum_{k=1}^{K}\prod_{i=0}^{1}\pi(\tau_i^*|\boldsymbol{y}_{(+1)}^T, \boldsymbol{D}, \boldsymbol{\alpha}^*, \mu^*, \boldsymbol{\theta}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{s}^{T(k)})\right]\\
+ \quad & \ln\left[K^{-1}\sum_{k=1}^{K}\prod_{i=0}^{1}\pi(\theta_i^*|\boldsymbol{y}_{(+1)}^T, \boldsymbol{D}, \boldsymbol{\alpha}^*, \mu^*, \boldsymbol{\tau}^*, \boldsymbol{\beta})^{(k)}, \boldsymbol{s}^{T(k)}\right]\\
+ \quad & \ln\left[K^{-1}\sum_{k=1}^{K}\pi(\boldsymbol{\beta}^*|\boldsymbol{y}_{(+1)}^T, \boldsymbol{D}, \boldsymbol{\alpha}^*, \mu^*, \boldsymbol{\tau}^*, \boldsymbol{s}^{T(k)})\right].
\end{aligned}
$$

This terms are obtained by running for K extra iterations of MCMC algorithm for each vector of parameters $\boldsymbol{\alpha}$, μ, $\boldsymbol{\tau}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, i.e running a total of 5K times the MCMC algorithm.

In the expression above all values labeled with superscript (k) are drawn from the full conditional posterior distributions of each parameter μ, $\boldsymbol{\tau}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. On the other hand $\boldsymbol{\alpha}_{i,j}^{(k)}$ with $i \neq j$ are drawn from the random walk in the following way:

$$\boldsymbol{\alpha}_{i,j}^{(k)} = \boldsymbol{\alpha}_{i,j}^* + U_A$$

where U_A is Multivariate Normal distributed with mean $\boldsymbol{\mu}_A$ and covariance matrix Σ_A. Additionally $q(\boldsymbol{\alpha}_{i,j}^*, \boldsymbol{\alpha}_{i,j}^{(k)})$ represents the multivariate Normal probability density function with mean $\boldsymbol{\alpha}_{i,j}^{(k)}$ and covariance matrix Σ_A measured in point $\boldsymbol{\alpha}_{i,j}^*$, and

$$A(\boldsymbol{\alpha}_{i,j}^{(k)}, \boldsymbol{\alpha}_{i,j}^*) = \min\left(1; \frac{\pi\left(\boldsymbol{\alpha}_{i,j}^{(k)}|\boldsymbol{s}^{T(k)}, \boldsymbol{Z}, \boldsymbol{\delta}\right)}{\pi\left(\boldsymbol{\alpha}_{i,j}^*|\boldsymbol{s}^{T(k)}, \boldsymbol{Z}, \boldsymbol{\delta}\right)}\right)$$

with $i \neq j$ is the acceptance ratio of $\boldsymbol{\alpha}_{i,j}^{(k)}$ in comparison with $\boldsymbol{\alpha}_{i,j}^*$ and $A(\boldsymbol{\alpha}_{i,j}^*, \boldsymbol{\alpha}_{i,j}^{(k)})$ its reciprocal.

For the homogenous case the first two terms of equation 4.9 are replaced by

$$\ln\left[K^{-1}\sum_{k=1}^{K}\prod_{i=0}^{1}\pi(\boldsymbol{\gamma}_i^*|\boldsymbol{y}_{(+1)}^T,\boldsymbol{D},\mu^{(k)},\boldsymbol{\tau}^{(k)},\boldsymbol{\theta}^{(k)},\boldsymbol{\beta}^{(k)},\boldsymbol{s}^{T(k)})\right]$$

and all the posterior distributions conditioned on $\boldsymbol{\alpha}^*$ are conditioned on $\boldsymbol{\gamma}_i^*$ with $i \in \{0,1\}$.

4.6 Results

4.6.1 Application to all data set

In a first phase all models were applied to the entire time series, from week 40/2008 to week 16/2011. This procedure had the objective of comparing models regarding the parameters estimates and the retrospective classification of each week in one of the Markov chain states (epidemic or non-epidemic), i.e. the decoding of the influenza activity states. Parameters and hidden states were estimated by a MCMC run of 200,000 iterations with a burn-in of 60,000 and a thinning of 100, for the non homogenous models, and 100,000 iterations with a burn-in of 25,000 and a thinning of 50, for the homogenous one. All results presented in this Chapter were obtained using specific programs implemented in the R computing language [100].

The MCMC output convergence was evaluated by the observation of the trace and auto-correlation functions of the parameters runs (Appendix C), and by the application of the statistic of Gelman-Rubin 1992 [101] and by the Raftery and Lewis method 1992 [102]. For this purpose the R package `coda` was used [103]. For all the parameters, the Gelman-Rubin scaling factors 97.5% percentiles were bellow 1.1, on the other hand the Raftery and Lewis method applied to the 1,500 runs after the burn-in and thin suggested for all the parameters a burn-in not superior to 2 and a maximum number of iterations close to 1,500. These results indicate that convergency has been achieved enabling the computations of the posterior means and credible intervals.

In Table 4.1 it can be seen that the non homogenous models present a better fit to data and, among these, model 2 does better.

According to Table 4.2 the estimates of parameters of the response variable are very similar between the three models. The only parameters that seem to change (decreasing in value) with the increase of model complexity, i.e. the introduction of

Model	ln(mL)
M0	-332.384
M1	-320.111
M2	-300.397

Table 4.1: Natural logarithm of the marginal likelihoods of the proposed models.

covariates to model the state transition probability matrix, are the parameters of the common part that has the objective of explaining ILI rate baseline behavior (μ, β_1 and β_2).

	Model 0		Model 1		Model 2	
Parameter	Mean	95% CI	Mean	95% CI	Mean	95% CI
τ_0	0.134	[0.091;0.187]	0.137	[0.093;0.189]	0.139	[0.091;0.196]
τ_1	0.006	[0.004;0.009]	0.006	[0.004;0.009]	0.006	[0.004;0.009]
μ	3.131	[2.191;4.120]	3.011	[1.995;3.993]	2.876	[1.845;3.950]
β_1	-0.861	[-1.660;-0.048]	-0.799	[-1.633;0.004]	-0.720	[-1.517;0.133]
β_2	1.947	[0.753;3.216]	1.861	[0.617;3.122]	1.604	[0.289;2.954]
θ_0	0.710	[0.526;0.879]	0.726	[0.554;0.888]	0.739	[0.571;0.902]
$\theta_{1,0}$	1.556	[1.353;1.759]	1.571	[1.375;1.777]	1.576	[1.375;1.779]
$\theta_{1,1}$	-0.005	[-0.007;-0.003]	-0.005	[-0.007;-0.003]	-0.005	[-0.007;-0.003]
$\gamma_{0,0}$	0.948	[0.888;0.986]	NA	NA	NA	NA
$\gamma_{0,1}$	0.052	[0.014;0.112]	NA	NA	NA	NA
$\gamma_{1,0}$	0.092	[0.025; 0.191]	NA	NA	NA	NA
$\gamma_{1,1}$	0.908	[0.809;0.975]	NA	NA	NA	NA
$\alpha_{0,1,0}$	NA	NA	-3.282	[-4.823;-1.953]	-3.267	[-5.077;-1.814]
$\alpha_{0,1,1}$	NA	NA	2.015	[-1.874;5.589]	1.995	[-1.917;5.917]
$\alpha_{0,1,2}$	NA	NA	NA	NA	0.778	[-4.822;5.626]
$\alpha_{1,0,0}$	NA	NA	-0.266	[-2.303;1.677]	0.326	[-1.943;2.786]
$\alpha_{1,0,1}$	NA	NA	-4.133	[-9.379;-0.826]	-1.320	[-7.438;3.698]
$\alpha_{1,0,2}$	NA	NA	NA	NA	-9.866	[-25.726;-0.192]

Table 4.2: Posteriori means and 95% credible intervals for model parameters. NA: not applicable.

Likewise the results obtained by Martinez-Beneito 2008 [25], for the homogenous model, a week is more likely to belong to the epidemic state if the previous week was in the epidemic state ($\gamma_{1,1}$ posterior mean of 0.91). This conservative result is also observed in the non-epidemic period, so a week is more likely to be in the non-epidemic state if the previous week was in the non-epidemic state ($\gamma_{0,0}$ posterior mean of 0.95). This result also means that in each week the mean posterior probabilities of entering in the epidemic state and also leaving the epidemic state are constant over time and very low (respectively 0.05 for $\gamma_{0,1}$ and 0.09 for $\gamma_{1,0}$).

For the non homogenous models (Figure 4.4), and as it was expected, the pos-

terior probabilities of changing the influenza activity state in each week vary over time. Nevertheless some of the posterior 95% credible intervals for the transition probabilities matrix parameters include zero (see Table 4.2). Exception are observed for the probability of entering the epidemic state, in the constant ($\alpha_{0,1,0}$), that is significant in both models, and for the probability of leaving the epidemic state, in the parameters associated with the early estimate of ILI rate ($\alpha_{1,0,1}$), in model 1, and the one associated with number of ILI cases positive for influenza in the previous week($\alpha_{1,0,2}$), in model 2.

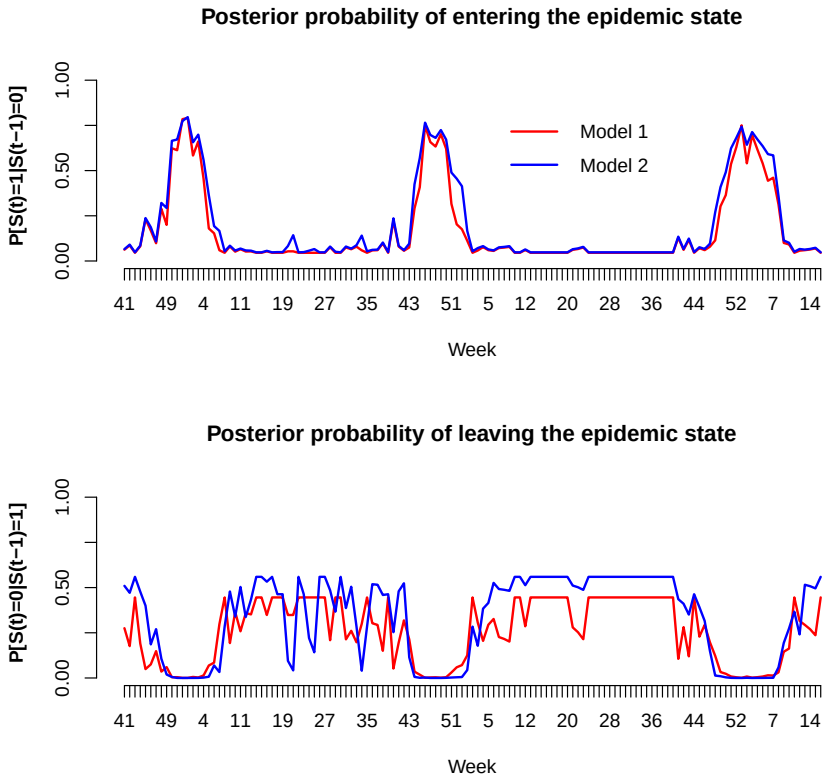

Figure 4.4: Mean posteriori probabilities of entering and leaving the epidemic influenza activity state according to the non-homogenous models (1 and 2).

Figure 4.5 depicts, for all the models, the weekly mean posterior probability of being in the epidemic sate. This value is concentrated in the neighborhood of zero or one, meaning that there is a low proportion of weeks with doubtful classification, i.e. with the posterior mean probability of being in the epidemic state near 0.5.

Concerning the classification of each week in the epidemic or in the non-epidemic state, a week was considered as epidemic (non epidemic) if the posteriori probability of belonging to the epidemic state was higher or equal to 0.5 (lower than 0.5). An epidemic period was defined as the consecutive set of weeks with the mean posterior probability of being in the epidemic state above 0.5.

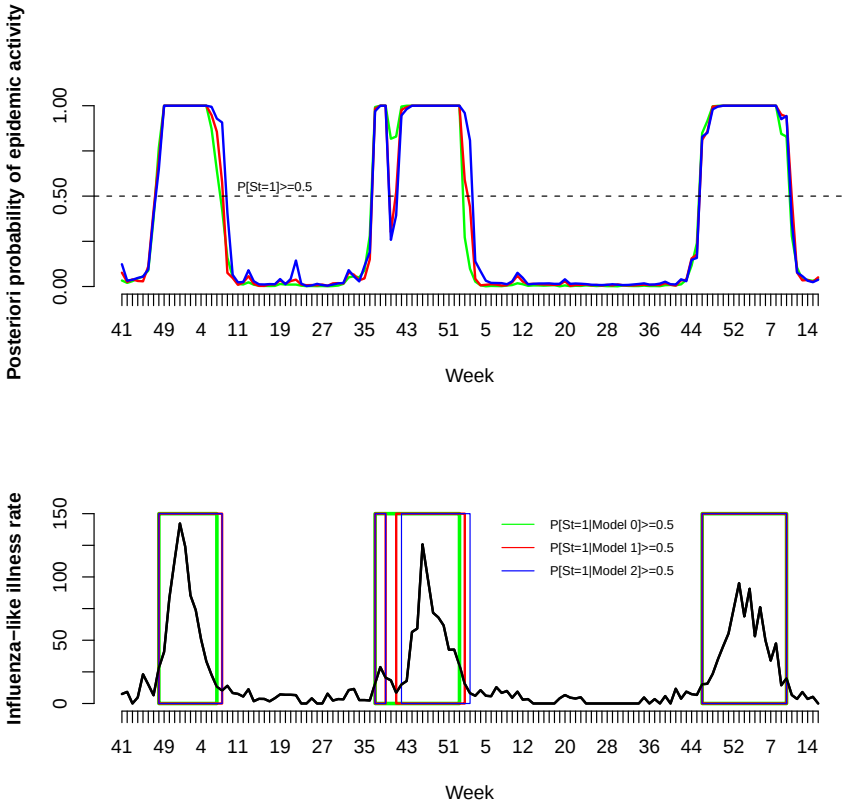

Figure 4.5: Panel 1: Mean posteriori probabilities of epidemic influenza activity (Model 0: green; Model 1: red; Model 2: blue). Panel 2 : Influenza-like illness rates, reported by Wednesday (solid line); periods of epidemic activity according to model fitted and probability threshold of influenza epidemic activity (colored boxes).

The non homogenous models identified four epidemic periods in the study period while the homogenous model identified three (Table 4.3). In Figure 4.5 it can be seen that in season 2009-2010, that corresponds to the (H1N1)2009 pandemic in the period from week 37/09 to week 2/10, the non homogenous models identified two distinct epidemics but the homogenous one considered only one epidemic period that started by week 37/09 and ended by week 53/09. On the other hand, the epidemic periods

for seasons 2008-09 and 2009-10 were consistently estimated by the three models.

Season	Model 0	Model 1	Model 2
2008-2009	48/08 to 7/09	48/08 to 8/09	47/08 to 8/09
2009-2010	37/09 to 53/10	37/09 to 39/09	37/09 to 39/09
		42/09 to 01/09	42/09 to 02/10
2010-2011	46/10 to 10/11	46/10 to 10/11	46/10 to 10/11

Table 4.3: Estimated influenza epidemic periods by proposed models for a posterior probability of being in the epidemic state higher than 0.5. Values represent week/year.

4.6.2 Real-time nowcast of 2010-11 influenza season

In Figure 4.6 the results of the weekly real-time influenza activity state nowcast and decoding are presented for season 2010-11. In the first panel one can see the mean posterior probability (mpp) of being in the epidemic state calculated in the same week using equation 4.5. The second panel shows the weekly mpp of being in the epidemic state calculated in the following week, i.e., with the ILI rate of the interest week totally observed. Finally the last panel presents the weekly mpp of being in the epidemic state calculated after all the ILI rate values for season 2010-11 are know, i.e. by week 17/2011.

At the end of the season, by week 17/2011, all the three models define the epidemic period from week 46/2010 to week 10/2011, presenting very similar epidemic state mpp. In general, regarding the nowcasting and the decoding of the last ILI rate observed, the non homogenous models presented an early increase and decrease in the weekly epidemic state mpp, specially model 2.

If one considers 0.5 as the mpp cut-off value to classify a week in the epidemic period, the first signal of the epidemic start is observed in week 48/2010 when model 2 decodes ILI rate of week 47/2010 as epidemic and nowcast for week 48/2010 a epidemic state mpp very close to 0.5. This tendency continues to be observed in week 49/2010, where model 2 calculates an epidemic state mpp in the neighborhood of 0.5 for the observed week 48/2010 ILI rate and nowcast for that week (49/2010) a epidemic state mpp clearly higher than 0.5. These results are then confirmed in week 50/2011 by all the three models after week 49/2010 ILI rate became known. It is important to underline that the non homogenous model 2 detects the signal of epidemic start two week before the homogenous model. On the other hand, looking for the signal of the end of the epidemic state, the non homogenous models also show

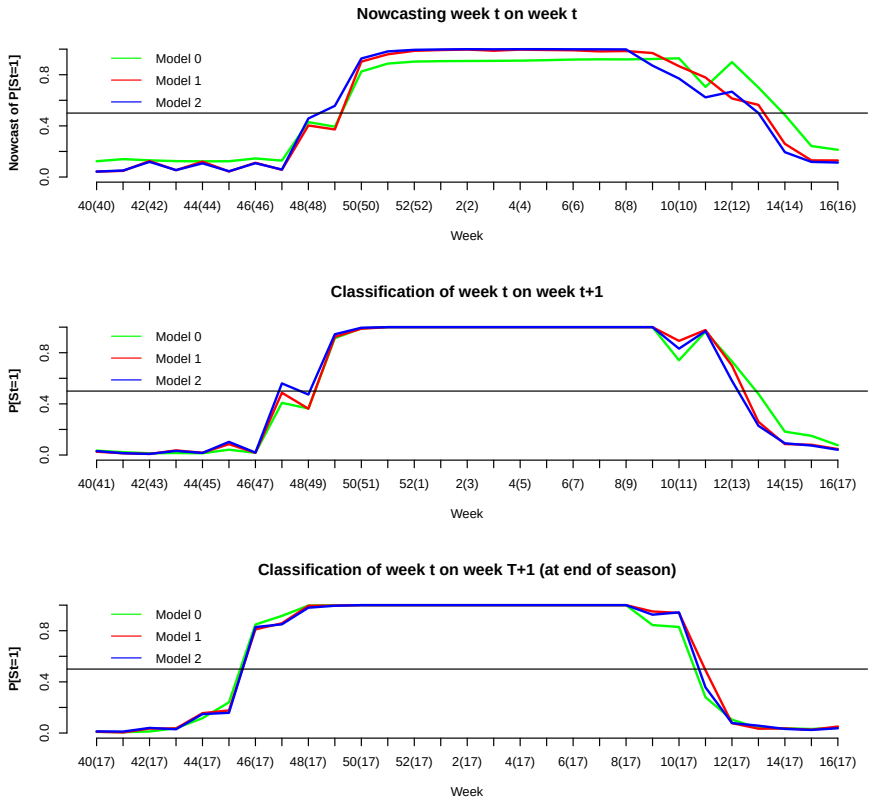

Figure 4.6: Weekly mean posteriori probabilities of epidemic influenza activity (season 2010-11) calculated Panel 1: in the current week (nowcast); Panel 2: in the following week; Panel 3: at end of the season. (.) week of the calculus.

better results, given that epidemic state mpp start to decrease earlier. Nevertheless it is only by week 13/2011 that model 2 nowcasts the end of the epidemic, what is confirmed in week 14/2011 when week 13/2011 ILI rate is finally known. The homogenous model is clearly more slowly in entering and leaving the epidemic state, which tends to increase the timeliness in detecting the epidemic start and end.

Furthermore the weekly ILI rates nowcasted by the three models (obtained by equation 4.6) do not present relevant differences, although non homogenous models present higher ILI rate estimates than the homogenous model. As can be seen in Figure 4.7, weekly ILI rates of season 2010-11 were predicted during the same week in a very satisfactory way, they start to increase, reached the peak and decrease in synchrony. This means that the models were able to tackle the ILI rate evolution by Friday of the same week, reducing reporting delay in 4 days.

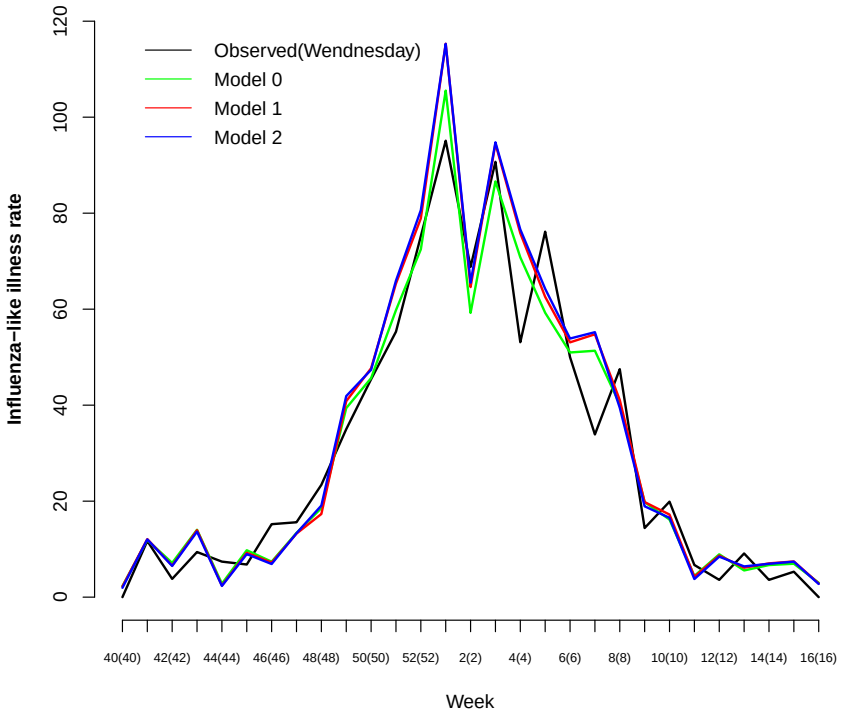

Figure 4.7: ILI rate nowcast for season 2010-11. (.) week of the calculus.

4.7 Discussion

Considering the main objective of this work, the models here proposed were able to nowcast the ILI incidence rate and the influenza activity state by Friday of the same week. These results were better achieved with the non homogenous HMM. As can be seen in the previous section, the non homogenous models were able to nowcast the beginning of the epidemic state two weeks before the homogenous model. On the other hand at the end of the epidemic, the probability of being in the epidemic state decreased more rapidly when calculated by the non homogenous models. Both these results underline the adding value of using non homogenous HMM to nowcast influenza epidemics. The inclusion of covariates, with early information on the epidemic evolution, to model the transition probabilities is of particulary importance to empower the model in the nowcasting task. Actually, using a homogenous HMM does not add too much to the nowcasting task given that the probability of a change

is immutable in time.

Also important to notice is the fact that when the models were fitted to the entire data set, the non homogenous models presented a better fit to the entire data set than the homogenous models, and identified in season 2009-10 (the pandemic (H1N1)2009) two distinct epidemic periods, where the homogenous model only identified one. During the 2009-10 influenza pandemic most of the European countries experienced two epidemic waves, one during Summer and a second one during Autumn and Portugal was an not exception. In fact, looking at other sources of information used during the pandemic (H1N1)2009 (e.g National Network of Laboratories for the A(H1N1) Diagnostic) a first small epidemic wave was identified in the period from week 31/2009 to week 39/2009 [107, 108]. All these features show that the non homogenous models give more reliable results.

Comparing both non homogenous models, model 1 and 2, model 2 presented the best results given that it produced the highest value of natural logarithm of the marginal likelihood and showed the higher capacity to nowcast the epidemic start in the real-time nowcast of season 2010-11.

Although these satisfactory results, the models proposed and data used have some caveats that must be addressed.

First the models were build using some ad-hoc decisions, more precisely regarding the structure of the state specific covariates used to model the response variable $y_{t(t)0}$ and $y_{t(t)1}$ and also when establishing, respectively a linear and quadratic relation between this estimates and the reported ILI rate $y_{t(t+1)}$. In the first decision we set the cut-offs at $v_0 = 1$ and $v_1 = 20$ based on the empirical knowledge of the surveillance system. To evaluate the importance of this decision we have also fitted to all the data set the homogenous model using other cut-off values for v_1, more specifically 5 and 10, without substantial changes in the results. Nevertheless, when applying these models to other surveillance systems, these fixed parameters should be tuned using empirical knowledge or exploratory data analysis.

Regarding the decision to set the association of the early estimate of the ILI rate as linear for the non epidemic state and quadratic for the epidemic state, other options, namely linear-linear, were also tested for the homogenous models presenting worst results, mainly in the real-time nowcast of the ILI rate.

Other point of interest, but that works against the non homogenous models, is the higher number of iterations, burn-in and thinning needed for the model MCMC

output to converge. This fact has particular impact in the computational time needed to obtain the nowcast results each week. The main reason for this is that the covariates transition probabilities parameters $\boldsymbol{\alpha}$ are sampled using the random-walk Metropolis-Hastings algorithm, and this algorithm has a slower convergence than the Gibbs algorithm. In this study we have used a logistic function to model these probabilities, which enable the direct sampling of $\boldsymbol{\alpha}_{0,1}$ and $\boldsymbol{\alpha}_{0,1}$ from the posterior full conditional distribution. A possible further development for the proposed model could be to choose a function to model the transition probabilities that could enable the use of the Gibbs algorithm to sample their specific parameters. Nevertheless this fact does not influence the timeliness of the nowcast objective, since the parameters estimation, that is more time consuming, can be obtain previously (by Wednesday) and the nowcast by Friday when the early estimate of the ILI rate is available.

Chapter 5

Conclusions

Generally the main research objectives of this thesis were all achieved with considerable success.

Regarding the **first objective**, *to unify in a single class the statistical methods, characterized by using interrupted mortality time series to estimated excess deaths attributable to influenza epidemics, in order to describe and compare their applicability and results*, the present work made possible to define a class of methods that contains the principal models, without covariates, that use interrupted time series to estimate influenza-associated deaths. They were categorized according to three main parameters: the model used to fit the interrupted time series and obtain a baseline, the a priori chosen type of periods used to estimate the influenza epidemic periods and the type of procedure used to fit the model to the time series (iterative or non iterative). This generalization led quite naturally to the construction of a set of user friendly `R`-routines, package `flubase`, for estimating these influenza-associated deaths with any of the methods in the class, that is available on the internet for download in `http://cran.r-project.org/web/packages/flubase/index.html` [27]. As it was developed, the `flubase` package can also be an important tool to compare the influenza associated deaths obtained by all the methods, just by varying each one of the parameters involved, allowing the researcher to perform sensitivity or specificity analysis of the obtained results.

Further, this package and methods framework has also shown to be useful to estimate excesses of deaths attributable to other events defined in time and with potential impact on the indicator of interest, e.g the excess deaths attributable to heat-waves [45, 46]. In this situation the E_a period is set as the period where the

heat-wave occurred.

The methods included in the framework were also applied to an example, the time series of weekly mortality due to P&I in Portugal for the period of 1980 to 2004. The results obtained suggest that:

1. In the absence of any reliable external source of information to define the epidemic periods (E_a) one should definitely choose the fixed period approach. Nonetheless the estimation of the E_a periods with more informative sources, leading to different E_a periods from year to year, is more desirable if one pretends a more conservative approach. This last option might be preferable given the evidence that: the duration and intensity of the influenza epidemics vary widely from season to season [35]; extracting all winter mortality data means that the baseline during these periods is estimated without any information from them; and also because the ecological nature of this type of studies requires precaution on the attribution of excess mortality to an event;

2. As it is known the seasonal ARIMA model is more adequate to model this type of time dependent data, producing non autocorrelated residuals and leading to a better estimate of the threshold, which is essential for the identification of the periods with excess deaths attributable to influenza (D_a). Given this, that has already been stated by others [13], one can ask why the ARIMA models are so rarely used. The answer might be that not only they are more difficult to understand and apply, but also the graphical aspect of the resulting baseline is not as smooth as the sinusoidal curve of the cyclic regression model. Actually, this visual characteristic has already been suggested to be a reason why government agencies (e.g US CDC) have chosen the regression approach [19] to estimate influenza associated deaths;

3. The iterative procedure should only be used for stationary time series, naturally leading to baselines further apart from the observed in those periods where a change in trend occurs, even with a very mild tendency as our application shows. When a serious lack of stationarity occurs the usual correction measures should be previously applied. Nonetheless it seems important to say that when the objective is to define a baseline for prospective identification and estimation of excess deaths in a mortality surveillance framework, this kind of methods are more useful [8, 9, 13, 48, 45].

The first objective did not include the study of the second group of methods used to estimate excess deaths attributable to influenza. This group is characterized by including in the model, to be adjusted to the time series mortality, an influenza activity indicator, like ILI incidence rates or number of influenza laboratory confirmed cases. Although some works have been presented where this models are compared with others [109], we considered that at this moment there is not yet a study that unify and describe this group of methods in a single class allowing a proper platform to compare them. This should be in our opinion one possible further work to be developed in this area.

Considering the **second objective**, *to estimate the excess mortality associated with the influenza epidemics occurred in Portugal in the period from 1980 to 2004 and compare the results with those from other locations*, the present study shows that in the period from 1980 to 2004, the seasonal average number of all cause excess deaths associated with influenza epidemics was 2,475 in Portugal, representing a crude excess all-cause death rate of 24.7 per 100,000 inhabitants. These seasonal influenza associated excess deaths ranged from 0, in five out of the 24 study seasons, to 8,514 in 1998-1999 season. Another important result was that in average 90% of the estimated excess deaths attributable to influenza occurred in people aged ≥ 65 years. All these results suggest that influenza epidemics in Portugal had in general the same profile as that described for other temperate countries in the Northern Hemisphere.

This is the first study to provide estimates of influenza seasonal mortality burden in the Portuguese population in a comprehensive way, an important step to set up references to contextualize the severity of the forthcoming influenza epidemics, but also for the design of rational national public health measures to mitigate influenza epidemics impacts.

A few issues could be further explored, in particular the potentially higher mortality burden in extreme age groups (0-4 and ≥ 65), compared to other countries. As in other countries, population ageing tends to increase the absolute burden of influenza, which has important consequences for disease control and public health strategies. Another important point to note is the fact that the impact of influenza epidemics in Portugal was relatively well captured in the all-cause mortality time series. This result supports the use of this indicator for influenza surveillance purposes in Portugal, in particular to give near real-time estimates of inter-pandemic and pandemic influenza mortality impact [45, 46, 5]. We note that all-cause mortality is less dependent on

diagnostic and coding differences between countries, time periods, or during unusual events like pandemics, as compared with cause-specific indicators. Further studies in other settings are warranted to confirm that all-cause mortality is an appropriate indicator to compare the impact of influenza epidemics at a regional and global scale [16, 110, 111].

Regarding the **third objective**, *to develop a statistical model to nowcast an influenza epidemic evolution* the work presented shows the advantage of using a non homogenous HMM to nowcast the ILI incidence rate and the influenza activity state in the context of a public health surveillance system. This advantage is achieved given that the non homogenous HMM enables the inclusion of covariates, with early information on the epidemic evolution, to model the influenza activity states transition probabilities. This feature has been shown to be of particular importance to empower the model in the nowcast task, mainly in comparison with the homogenous HMM, previously used in the literature, that does not add too much to the nowcasting task because the state transition probabilities are the same for all moments in time.

This thesis have also demonstrated that in Portuguese surveillance system the incomplete information from a GP based influenza surveillance enabled the early detection of the epidemic start. More specifically, it was possible to show that using a non homogenous HMM, with an early estimated of the ILI rate by Friday of the same week, improved the surveillance system timeliness in 2 weeks.

Since the proposed models were fitted to data that corresponds to three influenza seasons of the Portuguese ISS, and the real-time nowcast simulation was tested in one influenza season (2010-11), it is important to state that further applications of these models, to a higher number of seasons and also to data from other public health surveillance systems, are needed in order to warrant the adding value of using non homogenous HMMs to nowcast an epidemic evolution in the public health surveillance setting.

Bibliography

[1] Rebelo-de-Andrade H. Aspectos epidemiológicos e virológicos da gripe em Portugal. Desenvolvimento de um sistema de vigilância[dissertation]. [Lisboa]: Universidade de Lisboa;2001. 175 p.

[2] Fleming DM, Zambon M, Bartelds AIM, Jong JC de. The duration and magnitude of influenza epidemics: A study of surveillance data from sentinel general practices in England, Wales and the Netherlands. European Journal of Epidemiology. 1999; 15(5): 467-473.

[3] Simonsen L, Clark MJ, Schonberger, Arden N, Cox NJ, Fukuda K. Pandemic versus Epidemic Influenza Mortality: A Pattern of Changing Age Distribution. Journal of Infectious Diseases. 1998; 178(1): 53-60.

[4] Viboud C, Miller M, Olson D, Osterholm M, Simonsen L. Preliminary Estimates of Mortality and Years of Life Lost Associated with the 2009 A/H1N1 Pandemic in the US and Comparison with Past Influenza Seasons. PLoS Curr. 2010 Mar 20:RRN1153.

[5] Mazick A, Gergonne B, Wuillaume F, Danis K, Vantarakis A, et al. Higher all-cause mortality in children during autumn 2009 compared with the three previous years: pooled results from eight European countries. Euro Surveill: 2010; 15(5):pii=19480. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?Articled=1948

[6] European Centre for Prevention and Disease Control. European Influenza Surveillance Network [Internet]. Stockholm: European Centre for Prevention and Disease Control; [cited 2011 May 26]. Available from: http://www.ecdc.europa.eu/en/activities/surveillance/EISN/

[7] World Health Organization. WHO Global Influenza Surveillance Network [Internet]. Geneve: World Health Organization; [cited 2011 Jul 16]. Available from: http://www.who.int/csr/disease/influenza/surveillance/en/

[8] United States Centre For Disease Control and Prevention. Overview of Influenza Surveillance in the United States [Internet]. Atlanta: Centre for Disesase Control and Prevention; [cited 2011 Jul 16]. Available from: http://www.cdc.gov/flu/weekly/overview.htm

[9] EuroMOMO. European monitoring excess mortality for public health action [Internet]. Copenhagen: EuroMOMO; [cited 2011 Jul 16]. Available from: http://www.euromomo.eu/index.html

[10] Collins SD. Trend and Age Variation of Mortality and Morbidity from Influenza and Pneumonia. Public Health Monograph. 1957; 48.

[11] Viboud C, Boëlle PY, Carrat F, Valleron AJ, Flahault A. Prediction of the Spread of Influenza Epidemics by the Method of Analogues. American Journal of Epidemiology. 2003; 158(10): 996-1006.

[12] Donker T, van Boven M, van Ballegooijen WM, Van't Klooster TM, Wielders CC, Wallinga J. Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands. European Journal of Epidemiology. 2011;26(3):195-201. Epub 2011 Mar 18.

[13] Choi K and Thacker SB. An evaluation of influenza mortality surveillance 1962-1979. American Journal of Epidemiology. 1981; 113(3): 215-216.

[14] Lui K-J and Kendal AP. Impact pf influenza epidemics on mortality in the united states from october 1972 to may 1985. American Journal of Public Health. 1987; 77(6):712-716.

[15] Simonsen L, Clarke MJ, Williamson D, Stroup DF, Arden NH and Schonberger LB. The Impact of Influenza Epidemics on Mortality: Introducing a Severity Index. American Journal of Public Health. 1997; 87(12):1994-1950.

[16] Simonsen L, Reichert TA, Viboud C, Blackwelder WC, Taylor RJ, Miller MA. The impact of influenza vaccination on seasonal mortality in the US elderly population. Archives of Internal Medicine. 2005; 165(3)265-272.

[17] Zucs WHP, Buchholz U and Uphoff H. Influenza associated excess mortality in germany, 1985 - 2001. Emerging Themes in Epidemiology. 2005; 2(6):1-9.

[18] Rizzo C, Viboud C, Montomoli E, Simonsen L, Miller MA. Influenza-related mortality in the Italina elderly: No decline associated with the increasing vaccination coverage. Vaccine. 2006; 24(42-43):6468-6475.

[19] Antunes JL, Waldman EA, Borrell C, Paiva TM. Effectiveness of influenza vaccination and its impact on health inequalities. International Journal of Epidemiology. 2006; 36(6):1319-1326.

[20] Nunes B, Natário I, Carvalho ML. Time series methods for obtaining excess mortality attributable to influenza epidemics. Statistical Methods in Medical Research. 2011; 20(4):331-346. Epub 2010 March 8.

[21] Nogueira PJ, Rebelo-de-Andrade H. Excess mortality associated with influenza in Portugal (1991-1998). Poster session presented at: Option for Influenza Control IV. Creta, 2000.

[22] Nunes B, Viboud C, Machado A, Ringholz C, Rebelo-de-Andrade H, et al. (2011) Excess Mortality Associated with Influenza Epidemics in Portugal, 1980 to 2004. PLoS ONE 6(6): e20661. doi:10.1371/journal.pone.0020661

[23] Strat L, Carrat F. Monitoring epidemiologic surveillance data using Hidden Markov Chains models. Statistics in Medicine. 1999; 18(24); 3463-3478.

[24] Rath TM, Carreras M, Sebastiani P. Automated Detection of Influenza Epidemics. University of Massachusetts. 2003.

[25] Martínez-Beneito MA, Conesa D, Lopéz-Quiléz A, Lopez-Maside A. Bayesian Markov switching models for the early detection of influenza epidemics. Statistics in Medicine. 2008; 27(22); 4455-4468.

[26] Nunes B, Natário I, Carvalho ML. Nowcasting influenza epidemics using non homogenous hidden Markov models[Internet]. Notas e comunicações do Centro de Estatística e Aplicações da Universiade de Lisboa. 2011[cited August 2011];18. Available from:http://www.ceaul.fc.ul.pt/notas.html?ano=2011

[27] Nunes B, Natario I, Carvalho L. `flubase`: Baseline of mortality free of influenza epidemics. 2009: R package version 1.0.

[28] Cox NJ and Subbarao K. Global Epidemiology of Influenza: Past and Present. Annual Reviews of Medicine. 2000; 51:407-421.

[29] Clifford RE, Smith JWG, Tillet HE and Wherry PJ. Excess Mortality Associated with Influenza in England and Wales. International Journal of Epidemiology. 1977; 6(2):115-128.

[30] Alling DW, Blackwelder WC and Stuart-Harris CH. A Study of Excess Mortality During Influenza Epidemics in the United States, 1968–1976. American Journal of Epidemiology. 1981; 113(1):30-43.

[31] Tillet HE, Smith JWG and Gooch CD. Excess Deaths Attributable to Influenza in England and Wales: Age at Death and Certified Cause. International Journal of Epidemiology. 1983; 12(3):344-352.

[32] Sprenger MJ, Mulder PG, Meyer WE, Van Strik R and Masurel N. Impact of influenza on mortality in relation to age and underlying disease, 1967-1989. International Journal of Epidemiology. 1993; 22(2):334-340.

[33] Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, Anderson LJ and Fukuda K. Mortality Associated With Influenza and Respiratory Syncytial Virus in the United States. Journal of the American Medical Association. 2003; 289(2):179-186.

[34] Brinkhof MWG, Spoerri A, Birrer A, Hagman R, Koch D and Zwahlen M. Influenza-attributable mortality among the elderly in Switzerland. Swiss Medical Weekly. 2006; 136(2):302-309.

[35] Paget J, Marquet R, Meijer A, van der Velden K. Influenza activity in Europe during eight seasons (1999 to 2007): an evaluation of the indicators used to measure activity and an assessment of the timing, length and course of peak activity (spread) across Europe. BMC Infectious Diseases. 2007; **7**:141.

[36] Direcção-Geral da Saúde, Ministério da Saúde. Circular Informativa Nº35/DSCS/DPCD. Vacinação contra a gripe sazonal em 2007/2006. Available from: http://www.dgs.pt [3 June 2008].

[37] Barker WH, Borisute H and Cox C. A Study of the Impact of Influenza on the Functional Status of Frail Older People. Archives of Internal Medicine. 1998 Mar 23; 158(6):645-650.

[38] Doshi P. Are US flu death figures more PR than science? British Medical Journal. 2005;331:1412.

[39] Simonsen L, Taylor R, Viboud C, Dushoff J, and Miller MA. US flu mortality estimates are based on solid science. British Medical Journal 2006; 332:177-178.

[40] Box G and Jenkins G. Time series analysis: Forecasting and control. (1st edn) San Francisco: Holden-Day, 1970.

[41] Instituto Nacional de Saúde Dr. Ricardo Jorge. Gripe Pandémica e Sazonal: Programa de Intervençao do INSA[Pandemic and seasonal influeza: INSA intervention program]. Instituto Nacional de Saúde Dr. Ricardo Jorge; 2006.

[42] P. Nogueira and E. Paixão. Models for mortality associated with heatwaves: update of the Portuguese heat health warning system. International Journal of Climatology. 2007; 28 (4): 542-562.

[43] Hyndman RJ and Khandakar Y. Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software July 2008; 27:3.

[44] Calado R, Nogueira PJ, Catarino J, Paixão E, Botelho J, Carreira M, Falcão JM. A onda de calor de Agosto de 2003 e os seus efeitos sobre a população portuguesa. Revista Portuguesa de Saúde Pública. 2004;22(3):7-20.

[45] Nogueira PJ, Machado A, Rodrigues E, Nunes B, Sousa L, Jacinto M, Ferreira A, Falcão JM, Ferrinho P. The new automated daily mortality surveillance system in Portugal. Euro Surveillance. 2010;15(13):pii=19529. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19529

[46] Nogueira PJ, Nunes B , Machado A, Rodrigues E, Gómez V, Sousa L, Falcão J M. Early estimates of the excess mortality associated with the 2008-9 influenza season in Portugal. Euro Surveillance. 2009; 14 (18). Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19194

[47] Pelat C, Boëlle P-Y, Cowling BJ, Carrat F, Flahault A, Ansart S, and Valleron A-J. Online detection and quantification of epidemics. BMC Medical Informatics Decision Making. 2007; 7:29.

[48] Serfling RE. Methods for Current Statistical Analysis of Excess Mortality Pneumonia-Influenza Deaths. Public Heath Reports. 1963; 78 (6): 494-506.

[49] Reichert TA, Simonsen L, Sharma A, Pardo SA, Fedson DS, et al. Influenza and the winter increase in mortality in the United States, 1959-1999. American Journal of Epidemiology. 2004; 160(5): 492-502.

[50] Simonsen L, Fukuda K, Schonberger LB, Cox NJ. The impact of influenza epidemics on hospitalizations. Journal of Infectious Diseases.2000; 181(3): 831-837.

[51] Baltussen R, Reinders A, Sprenger MJW, Postma MJ, Jager JC, et al. Estimating influenza-related hospitalization in the Netherlands. Epidemiology and Infection. 1998; 121(1): 129-138.

[52] Francisco P, Donalisio MRD, Lattore MDD. Impact of influenza vaccination on mortality by respiratory diseases among Brazilian elderly persons. Revista De Saude Publica. 2005; 39(1): 75-81.

[53] Oropesa S, Acosta B, Pinon A, Andreus H, Hernandez B, et al. The impact of influenza vaccination in the reduction of morbidity and in the exacerbation in asthmatic patients. Options for the Control of Influenza V; 2004; 1263: 355-359.

[54] Kyncl J, Prochazka B, Goddard NL, Havlickova M, Castkova J, et al. A study of excess mortality during influenza epidemics in the Czech Republic, 1982-2000. European Journal of Epidemiology. 2005; 20(4): 365-371.

[55] Centre for Disease Control and Prevention. UEstimates of Deaths Associated with Seasonal Influenza - United States, 1976-2007. MMWR 2010; 59 (33): 1057-1062.

[56] Meier CR, Napalkov PN, Wegmuller Y, Jefferson T, Jick H. Population-based study on incidence, risk factors, clinical complications and drug utilisation associated with influenza in the United Kingdom. European Journal of Clinical Microbiology & Infectious Diseases. 2000; 19(11): 834-842.

[57] Thompson WW, Shay DK, Weintraub E, Brammer I, Bridges CB, et al. Influenza-associated hospitalizations in the United States. Journal of the American Medical Association. 2004; 292(11): 1333-1340.

[58] Schanzer DL, Tam TWS, Langlev JM, Winchester BT. Influenza-attributable deaths, Canada 1990-1999. Epidemiology and Infection. 2007; 135(7): 1109-1116.

[59] Yap FHY, Ho PL, Lam KF, Chan PKS, Cheng YH, et al. Excess hospital admissions for pneumonia, chronic obstructive pulmonary disease, and heart failure during influenza seasons in Hong Kong. Journal of Medical Virology. 2004; 73(4): 617-623.

[60] Instituto Nacional de Estatística. Infoline[Internet]. Instituto Nacional de Estatística. 2008 [cited in 2009]. Available from http:
www.ine.pt

[61] Associação Portuguesa de Médicos de Clínica Geral. IPCP-2 Classificação Internacional de Cuidados Primários[ICPC-2 International Classification of Primary Care, second edition. WONCA international Classification Committee. Portuguese version edited by the Portuguese Association of General Practitioners]. Lisboa. Associação Portuguesa de Médicos de Clínica Geral; 1999.

[62] Instituto Nacional de Saúde Dr. Ricardo Jorge. Médicos-Sentinela. O que se fez em 2007.[Portuguese GP Network. What was done in 2007]. Lisboa. Instituto Nacional de Saúde Dr. Ricardo Jorge; 2009.

[63] United Nations, Population Division, Population Estimates and Projections Section. World Population Prospects: the 2008 revision. United Nations, Department of Economics and Social Affairs, Population Division, Population Estimates and Projections Section; 2008.

[64] Fenton LF. The sum of log-normal probability distributions in scatter transmission systems. IRE Trans Commun: 1960; 8.

[65] Beigel J. Influenza. Critical Care Medicine: 2008 36: 2660-2666.

[66] Branco MJ, Nogueira P. De que mais se morre em Portugal [Main causes of death in Porugal]. Lisboa: Instituto Nacional de Saúde Dr. Ricardo Jorge; 2003. 162 p.

[67] World Health Organization. Mortality Country Fact Sheet 2006 - Canada. Geneve: WHO; 2004.

[68] Madjid M, Aboshady I, Awan I, Litovsky S, Casscells S. Influenza and cardiovascular disease: is there a causal relationship? Texas Heart Institute Jounal. 2004; 31(1): 4-13.

[69] Nunes B, Marinho Falcão J.Vacina antigripal: cobertura da população portuguesa entre 1998/1999 a 2007/2008. [Influenza vaccine: coverage of the portugese population from 1998/1999 to 2007/2008] Lisboa: Instituto Nacional de Saúde Dr. Ricardo Jorge (INSA); 2008.

[70] Jansen A, Sanders EAM, Hoes AW, van Loon AM, Hak E. Influenza- and respiratory syncytial virus-associated mortality and hospitalisations. European Respiratory Journal. 2007; 30(6): 1158-1166.

[71] Zambon MC, Stockton JD, Clewley JP, Fleming DM. Contribution of influenza and respiratory syncytial virus to community cases of influenza-like illness: an observational study. Lancet. 2001; 358(9291): 1410-1416.

[72] Iwane MK, Edwards KM, Szilagyi PG, Walker FJ, Griffin MR, et al. Population-based surveillance for hospitalizations associated with respiratory syncytial virus, influenza virus, and parainfluenza viruses among young children. Pediatrics. 2004; 113(6): 1758-1764.

[73] Shay DK, Holman RC, Newman RD, Liu LL, Stout JW, et al. Bronchiolitis-associated hospitalizations among US children, 1980-1996. Journal of the American Medical Association.1999(15); 282: 1440-1446.

[74] Centre for Disease Control and Prevention. Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group.MMWR 2001; 50 (RR13): 1-35

[75] Centre for Disease Control and Prevention. Framework for evaluating public health surveillance systems for early detection of outbreaks.MMWR 2004; 53 (RR05): 1-11

[76] Lombardo JS, Ross D. Disease Surveillance, a Public Health Priority. In:Lombardo JS, Buckeridge DL, editors. Disease Surveillance. Hoboken, New Jersey: Jonh Wiley and Sons; 2007. p. 1-35.

[77] Fleming DM, van der Velden J and Paget J.The evolution of influenza surveillance in Europe and prospects for the next 10 years. Vaccine. 2003; 21(16): 1749-1753.

[78] Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. Journal of the Royal Statistics Society Series A. 2003; 166(1): 5-21.

[79] Farrington CP, Andrews N. Outbreak detection: Applications to infectious disease surveillance. In: Brookmeyer R, Stroup DF, editors. Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance. Oxford: Oxford University Press; 2004. p. 203-231.

[80] Buckridge DL, Burkom HS, Campell M, Hogan WR, Moore A. Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics. 2005; 38(2): 99-113.

[81] Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review[Internet]. Milton Keynes, UK: The Open University; 2010 [cited 2011 Jun]. 53 p. Available from:http://stats-www.open.ac.uk/TechnicalReports/OutbreakReviewPaper.pdf

[82] Hohle M. surveillance: An R package for the monitoring of infectious diseases. Computational Statistics. 2007; 22(4): 37-47.

[83] Deckers JG, Paget WJ, Schellevis FG, Fleming DM. European primary care surveillance networks: their structure and operation. Fam Pract. 2006 Apr;23(2):151-8.

[84] Fleming DM and Elliot AJ. Lessons from 40 years' surveillance of influenza in England and Wales. Epidemiology Infection. 2008 July; 136(7): 866-875.

[85] Truyers C, Lesaffre E, Bartholomeeusen S, Aertgeerts B, Snacken R, Brochier B, Yane F, Buntinx F. Computerized general practice based networks yield comparable performance with sentinel data in monitoring epidemiological time-course of influenza-like illness and acute respiratory illness. BMC Fam Pract. 2010 Mar 22;11:24.

[86] Nicoll A, Ammon A, Amato Gauci A, Ciancio B, Zucs P, Devaux I, Plata F, Mazick A, Mølbak K, Asikainen T, Kramarz P. Experience and lessons from surveillance and studies of the 2009 pandemic in Europe. Public Health. 2010; 124(1):14-23.

[87] Baldi P, Chauvin Y, Hunkapillar T, McClure MA. Hidden Markov Models of biological primary sequences information. Proceedings of the National Academy of Sciencies. 1994; 91:1059-1063.

[88] Leroux BG. Maximum-likelihood estimation for hidden Markov models. Statistics Processes and their Applications. 1992; 40:127-143.

[89] Wang P, Puterman M. Analysis of longitudinal data of epileptic seizure counts-A two state hidden Markov regression approach. Biometrical Journal. 2001; 43(8):941-962.

[90] Bureau A, Shiboski S, Hughes JP. Application of continous time hidden Markov models to the study of misclassified diseases outcomes. Statistics in Medicine. 2003; 22(3):441-462.

[91] Cooper B, Lipsitch M. The analysis of hospital infection data using hidden Markov models. Biostatistics. 2004;5(2); 223-237.

[92] Watkins RE, Eagleson S, Veenendaal B, Wright G, Plant AJ. Disease surveillance using hidden Markov models. BMC Medical informatics and Decision making. 2009; 9:39.

[93] Filardo AJ, Gordon SF. Business cycle durations. Jornal of Econometrics. 1998; 85:99-123.

[94] Hughes JP, Guttorp P, Charles SP. A Non-Homogeneous Hidden Markov Model for Precipitation Occurrence. Applied Statistics. 1999; 48 (1):15-30.

[95] Paroli R, Spezia L. Bayesian inference in non-homogeneous Markov mixtures of periodic autoregressions with state-dependent exogenous variables. Computationals Statistics and Data Analysis. 2008; 52(5):2311-2330.

[96] Smith T, Vounatsou P. Estimation of infection and recovery rates for highly polymorohic parasites when dectability is imperfect, using hidden Markov models. Statistics in Medicine. 2002; 22(10):1709-1724.

[97] Zucchini W, MacDonald IL. Hidden Markov Models for Time Series: An introduction using R. Boca Raton: Chapman and Hall/CRC; 2009 275p.

[98] Falcão IM, de Andrade HR, Santos AS, Paixão MT, Falcão JM. Programme for the surveillance of influenza in Portugal: results of the period 1990-1996. J Epidemiol Community Health. 1998 Apr;52 Suppl 1:39S-42S.

[99] Chib S. Calculating posterior distributions and modal estimates in Markov mixtures models. Journal of Econometrics. 1996 Apr;75:79-97.

[100] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2009. ISBN 3-900051-07-0, URL http://www.R-project.org.

[101] Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences (with discussion). Statistical Science. 1992; 7:457-511.

[102] Raftery AE, Lewis S. How Many Iterations in the Gibbs Sampler? In: Bernardo JM, Berger J, Dawid AP, Smith AMF, editors. Bayesian statistics. Oxford: Oxford University Press; 1992. p. 763-773.

[103] Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 2006; 6(1):7-11.

[104] Chib S. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association. 1995;90 (432): 1313-1321.

[105] Chib S, Jeliazkov I. Marginal Likelihood From the Metropolis-Hastings Output. Journal of the American Statistical Association. 2001;96 (453): 270-281.

[106] Neal RM. Erronous results in "Marginal likelihood from the Gibbs output". Unpublished manuscript.[internet]. Available from: http://www.cs.utoronto.ca/radford/

[107] Flasche S, Hens N, Boelle PY, Mossong J, van Ballegooijen WM, Nunes B, Rizzo C, Popovici F, Santa-Olalla P, Hruba F, Parmakova K, Baguelin M, van Hoek AJ, Desenclos JC, Bernillon P, Camara AL, Wallinga J, Asikainen T, White PJ, Edmunds WJ. Different transmission patterns in the early stages of the influenza A(H1N1)v pandemic: A comparative analysis of 12 European countries. Epidemics. 2011 Jun;3(2):125-33. Epub 2011 Apr 13.

[108] Departamento de Doenças Infecciosas e Departamento de Epidemiologia. Instituto Nacional de Saude Dr. Ricardo Jorge. Relatório do Programa Nacional da Vigilância da Gripe - epocas 2008-09 e 2009-10 [Influenza Surveillance National Program Report - seasons 2008-09 and 2009-10]. 2010 [Internet]. Lisboa: 2010. Instituto Nacional de Saude Dr. Ricardo Jorge. [cited 2011 Jul]. 61 p. Available from:http://www.insa.pt/sites/INSA/Portugues/Publicacoes/Outros/Documents/DoencasInfecciosas/.

[109] Thompson W, Weintraub E, Dhankhar P, Cheng P-Y, Brammer L, Meltzer M, Bresee J, Shay D. Estimates of US influenza-associated deaths made using four different methods Influenza and Other Respiratory Viruses. 2009;3 (1): 37-49.

[110] Mazick A, Europe Wommi. Monitoring excess mortality for public health action: potential for a future European network. Euro Surveillance. 2007; 12(1):pii=3107. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=3107

[111] Richard SA, Sugaya N, Simonsen L, Miller MA, Viboud C. A comparative study of the 1918-1920 influenza pandemic in Japan, USA and UK: mortality impact and implications for pandemic planning. Epidemiology and Infection. 2009: 137(8); 1062-1072.

Appendix A

SARIMA best-fitted models

Outcome	Model	Box-Ljung test for auto-correlation of residuals
All causes		
0-4	$ARIMA(0,1,2)(1,0,1)_{12}$	$X = 2.5754, df = 5.663, p = 0.8321$
5-54	$ARIMA(2,1,1)(2,0,2)_{12}$	$X = 3.2763, df = 5.663, p = 0.7379$
55-64	$ARIMA(2,1,2)(1,0,1)_{12}$	$X = 4.3128, df = 5.663, p = 0.5929$
65-69	$ARIMA(1,1,2)(2,0,1)_{12}$	$X = 3.4083, df = 5.663, p = 0.7194$
70-74	$ARIMA(4,1,2)(2,0,1)_{12}$	$X = 1.1144, df = 5.663, p = 0.9739$
75-79	$ARIMA(2,1,1)(2,0,0)_{12}$	$X = 6.3192, df = 5.663, p = 0.34$
80-84	$ARIMA(1,1,2)(2,0,1)_{12}$	$X = 2.9072, df = 5.663, p = 0.7886$
≥ 85	$ARIMA(1,1,1)(2,0,1)_{12}$	$X = 6.7818, df = 5.663, p = 0.3053$
Diseases of respiratory system		
0-4	$ARIMA(1,1,2)(1,0,1)_{12}$	$X = 0.201, df = 5.663, p = 0.9997$
5-54	$ARIMA(2,1,1)(0,0,2)_{12}$	$X = 4.3446, df = 5.663, p = 0.5885$
55-64	$ARIMA(0,1,3)(0,0,2)_{12}$	$X = 8.1654, df = 5.663, p = 0.1979$
65-69	$ARIMA(1,1,1)(2,0,0)_{12}$	$X = 4.356, df = 5.663, p = 0.5869$
70-74	$ARIMA(2,1,1)(2,0,1)_{12}$	$X = 1.5089, df = 5.663, p = 0.9465$
75-79	$ARIMA(1,0,0)(2,0,2)_{12}$	$X = 2.429, df = 5.663, p = 0.8504$
80-84	$ARIMA(1,1,1)(1,0,1)_{12}$	$X = 2.5779, df = 5.663, p = 0.8318$
≥ 85	$ARIMA(5,1,3)(2,0,1)_{12}$	$X = 0.7527, df = 5.663, p = 0.9902$
Chronic respiratory diseases		
0-4	$ARIMA(0,1,1)(1,0,0)_{12}$	$X = 8.7101, df = 5.663, p = 0.1654$
5-54	$ARIMA(1,1,1)(0,0,2)_{12}$	$X = 7.2876, df = 5.663, p = 0.26$
55-64	$ARIMA(2,1,2)(1,0,1)_{12}$	$X = 4.8272, df = 5.663, p = 0.5239$
65-69	$ARIMA(0,1,4)(0,0,2)_{12}$	$X = 5.4006, df = 5.663, p = 0.451$
70-74	$ARIMA(1,1,0)(2,0,1)_{12}$	$X = 29.0582, df = 5.663, p < 0.001$
75-79	$ARIMA(4,1,1)(2,0,1)_{12}$	$X = 4.0111, df = 5.663, p = 0.6347$
80-84	$ARIMA(2,1,1)(2,0,1)_{12}$	$X = 2.4044, df = 5.663, p = 0.8535$
≥ 85	$ARIMA(3,0,4)(1,0,1)_{12}$	$X = 2.917, df = 5.663, p = 0.787$
Pneumonia and Influenza		
0-4	$ARIMA(3,1,3)$	$X = 1.6593, df = 5.663, p = 0.9336$
5-54	$ARIMA(2,0,1)(0,0,1)_{12}$	$X = 3.4964, df = 5.663, p = 0.707$
55-64	$ARIMA(2,0,1)(1,0,1)_{12}$	$X = 6.496, df = 5.663, p = 0.332$
65-69	$ARIMA(2,0,1)(2,0,2)_{12}$	$X = 4.8612, df = 5.663, p = 0.5194$
70-74	$ARIMA(3,0,3)(2,0,2)_{12}$	$X = 1.7874, df = 5.663, p = 0.9216$
75-79	$ARIMA(1,0,2)(2,0,2)_{12}$	$X = 1.7948, df = 5.663, p = 0.9209$
80-84	$ARIMA(2,1,1)(2,0,2)_{12}$	$X = 2.6166, df = 5.663, p = 0.8268$
≥ 85	$ARIMA(0,1,1)(1,0,1)_{12}$	$X = 1.2253, df = 5.663, p = 0.9672$
Cardiovascular disease		
0-4	$ARIMA(1,0,1)(1,0,1)_{12}$	$X = 9.8662, df = 5.663, p = 0.1114$
5-54	$ARIMA(0,1,2)(1,0,2)_{12}$	$X = 5.4022, df = 5.663, p = 0.4516$
55-64	$ARIMA(2,1,2)(1,0,1)_{12}$	$X = 1.003, df = 5.663, p = 0.9798$
65-69	$ARIMA(1,1,1)(2,0,1)_{12}$	$X = 3.9818, df = 5.663, p = 0.6388$
70-74	$ARIMA(3,1,1)(2,0,2)_{12}$	$X = 1.0118, df = 5.663, p = 0.9794$
75-79	$ARIMA(3,1,3)(1,0,1)_{12}$	$X = 0.1189, df = 5.663, p = 1$
80-84	$ARIMA(2,1,1)(2,0,2)_{12}$	$X = 4.2804, df = 5.663, p = 0.5973$
≥ 85	$ARIMA(1,1,2)(2,0,0)_{12}$	$X = 0.3565, df = 5.663, p = 0.9986$
Ischemic heart disease		
0-4	$ARIMA(0,1,1)(1,0,0)_{12}$	$X = 2.3994, df = 5.663, p = 0.854$
5-54	$ARIMA(1,1,2)(2,0,1)_{12}$	$X = 1.236, df = 5.663, p = 0.9665$
55-64	$ARIMA(2,1,3)(1,0,2)_{12}$	$X = 1.5871, df = 5.663, p = 0.94$
65-69	$ARIMA(4,1,3)(1,0,1)_{12}$	$X = 1.4818, df = 5.663, p = 0.9487$
70-74	$ARIMA(2,1,1)(2,0,2)_{12}$	$X = 0.9129, df = 5.663, p = 0.984$
75-79	$ARIMA(2,1,1)(1,0,2)_{12}$	$X = 5.2412, df = 5.663, p = 0.4713$
80-84	$ARIMA(1,1,0)(1,0,1)_{12}$	$X = 35.1582, df = 5.663, p < 0.001$
≥ 85	$ARIMA(2,0,4)(0,0,1)_{12}$	$X = 1.343, df = 5.663, p = 0.9592$

Table A.1: Seasonal ARIMA best-fitted models by R package forecast and Box-Ljong test for residuals auto correlation

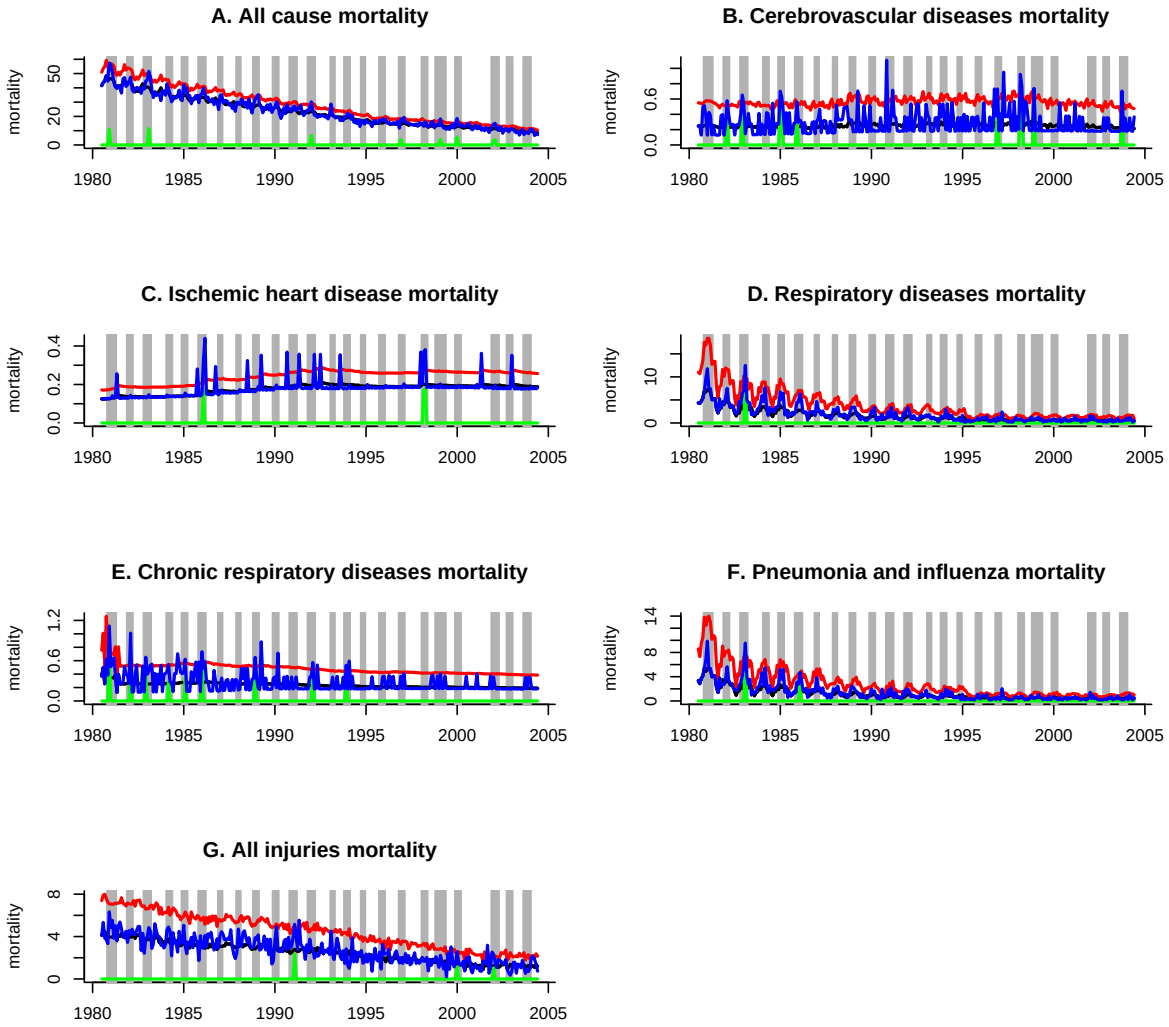

Figure A.1: Mortality rates (blue), mortality baseline (black) and 95% confidence limit (red), estimated excess death rate (green) by month and influenza epidemic periods (grey rectangles) for the study causes of death - age group 0-4

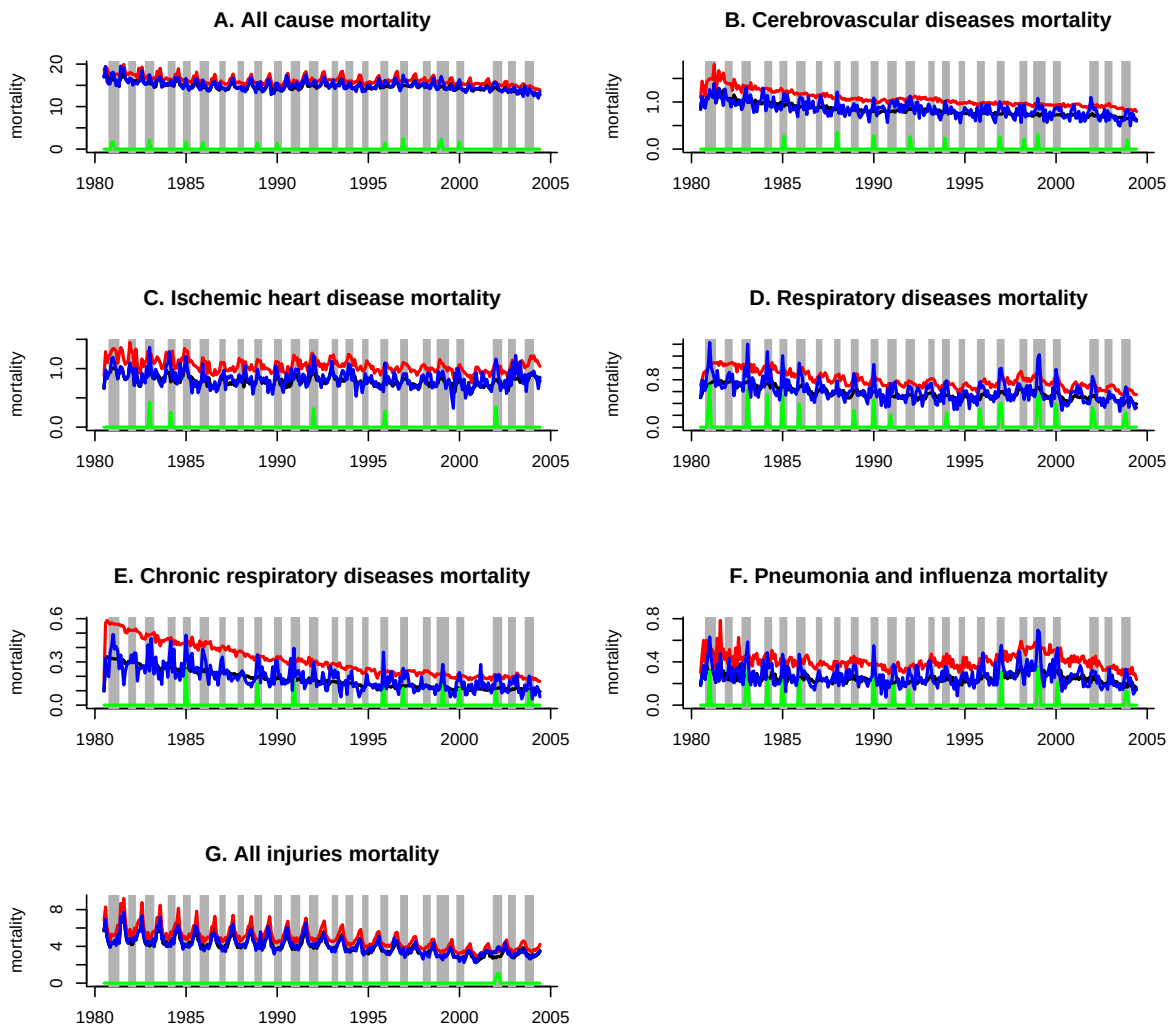

Figure A.2: Mortality rates (blue), mortality baseline (black) and 95% confidence limit (red), estimated excess death rate (green) by month and influenza epidemic periods (grey rectangles) for the study causes of death - age group 5-54

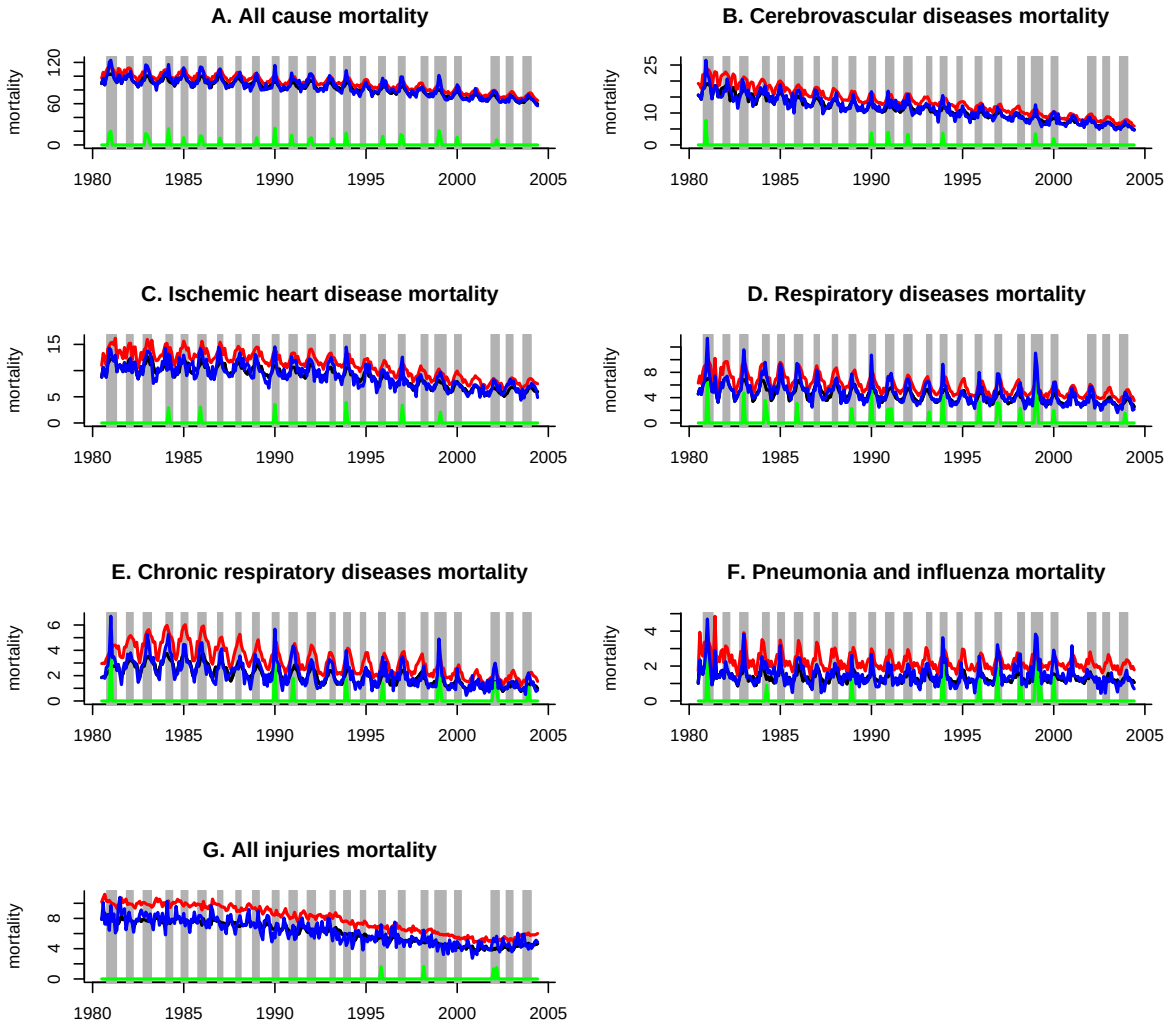

Figure A.3: Mortality rates (blue), mortality baseline (black) and 95% confidence limit (red), estimated excess death rate (green) by month and influenza epidemic periods (grey rectangles) for the study causes of death - age group 55-64

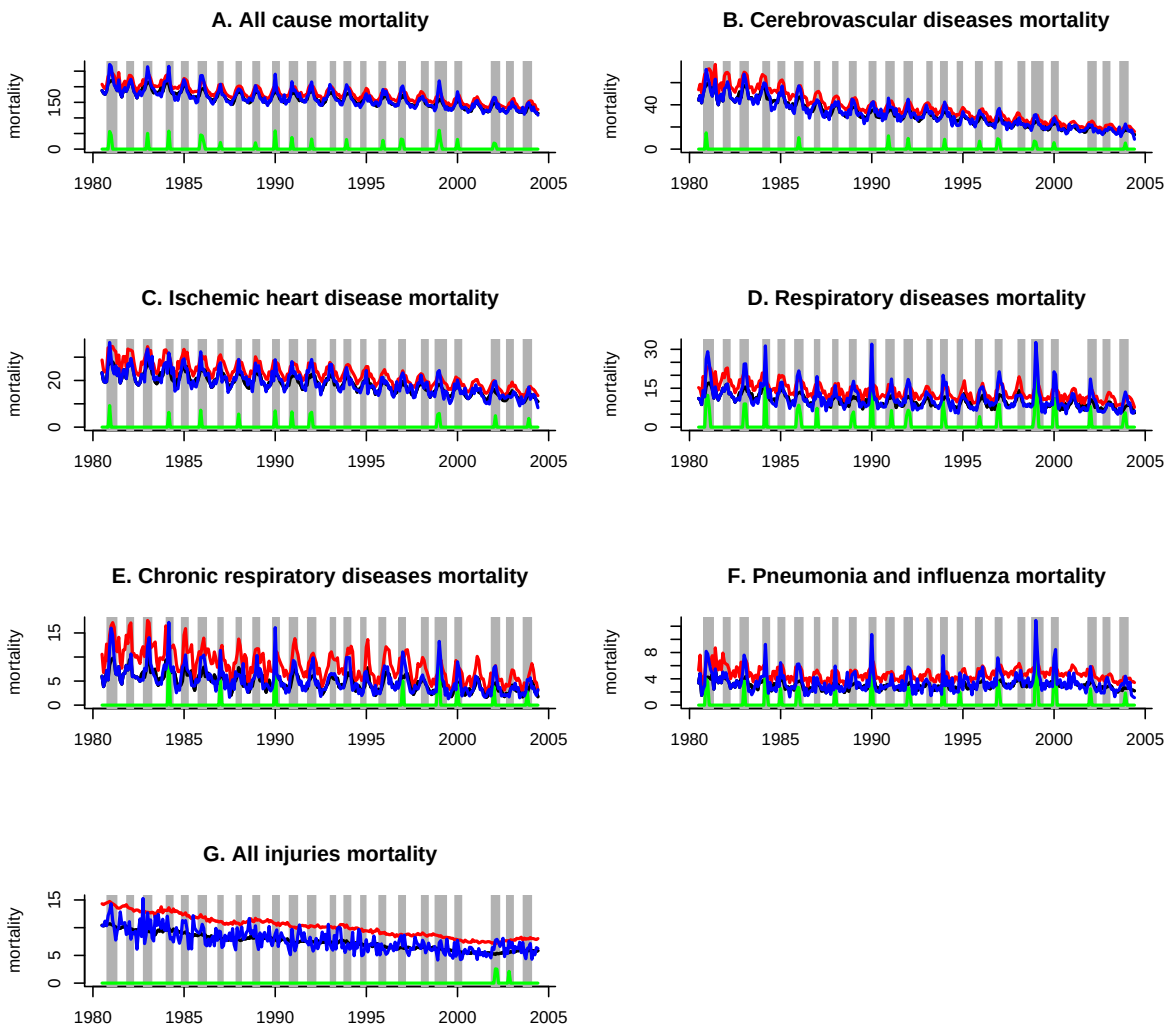

Figure A.4: Mortality rates (blue), mortality baseline (black) and 95% confidence limit (red), estimated excess death rate (green) by month and influenza epidemic periods (grey rectangles) for the study causes of death - age group 65-69

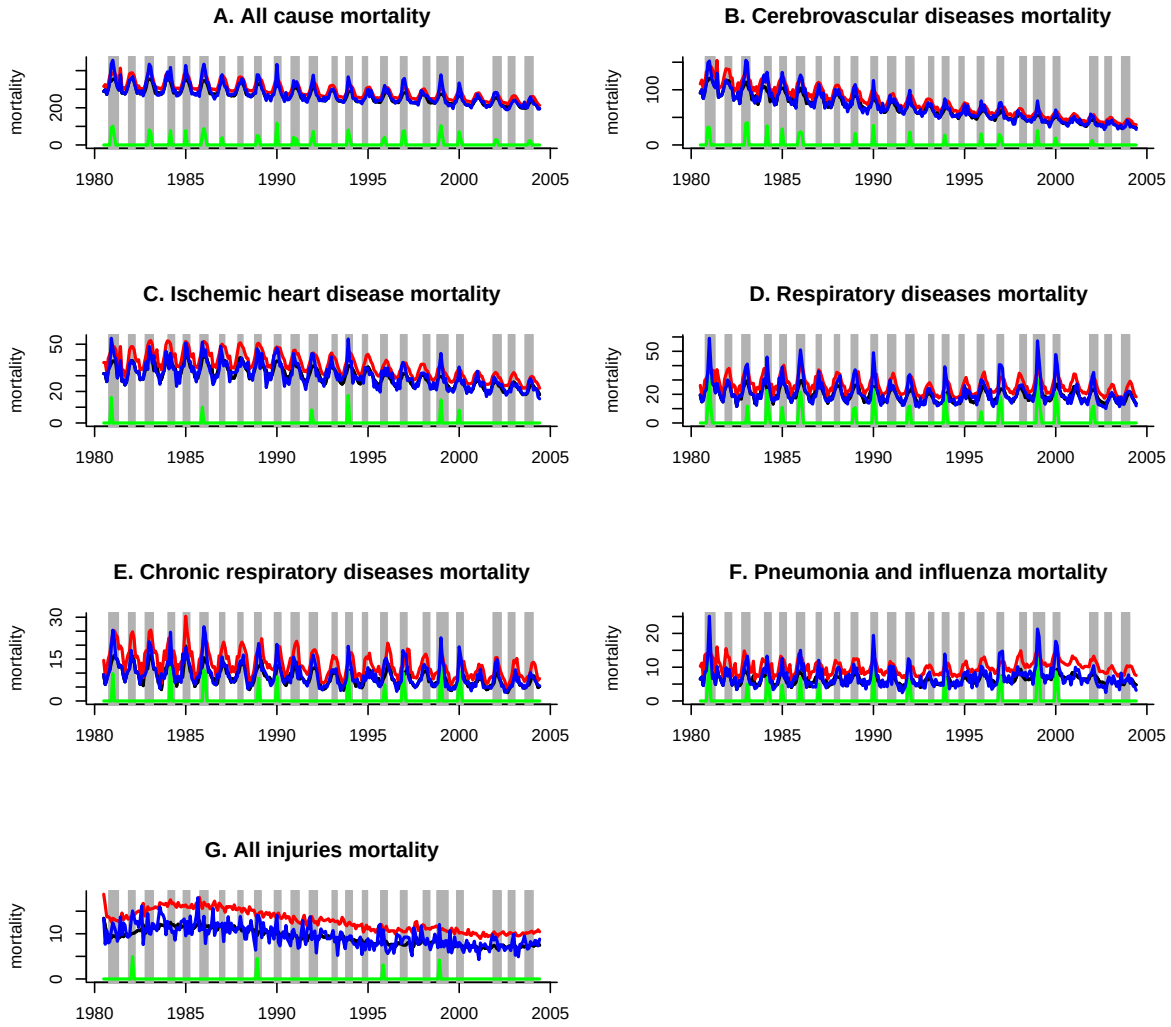

Figure A.5: Mortality rates (blue), mortality baseline (black) and 95% confidence limit (red), estimated excess death rate (green) by month and influenza epidemic periods (grey rectangles) for the study causes of death - age group 70-74

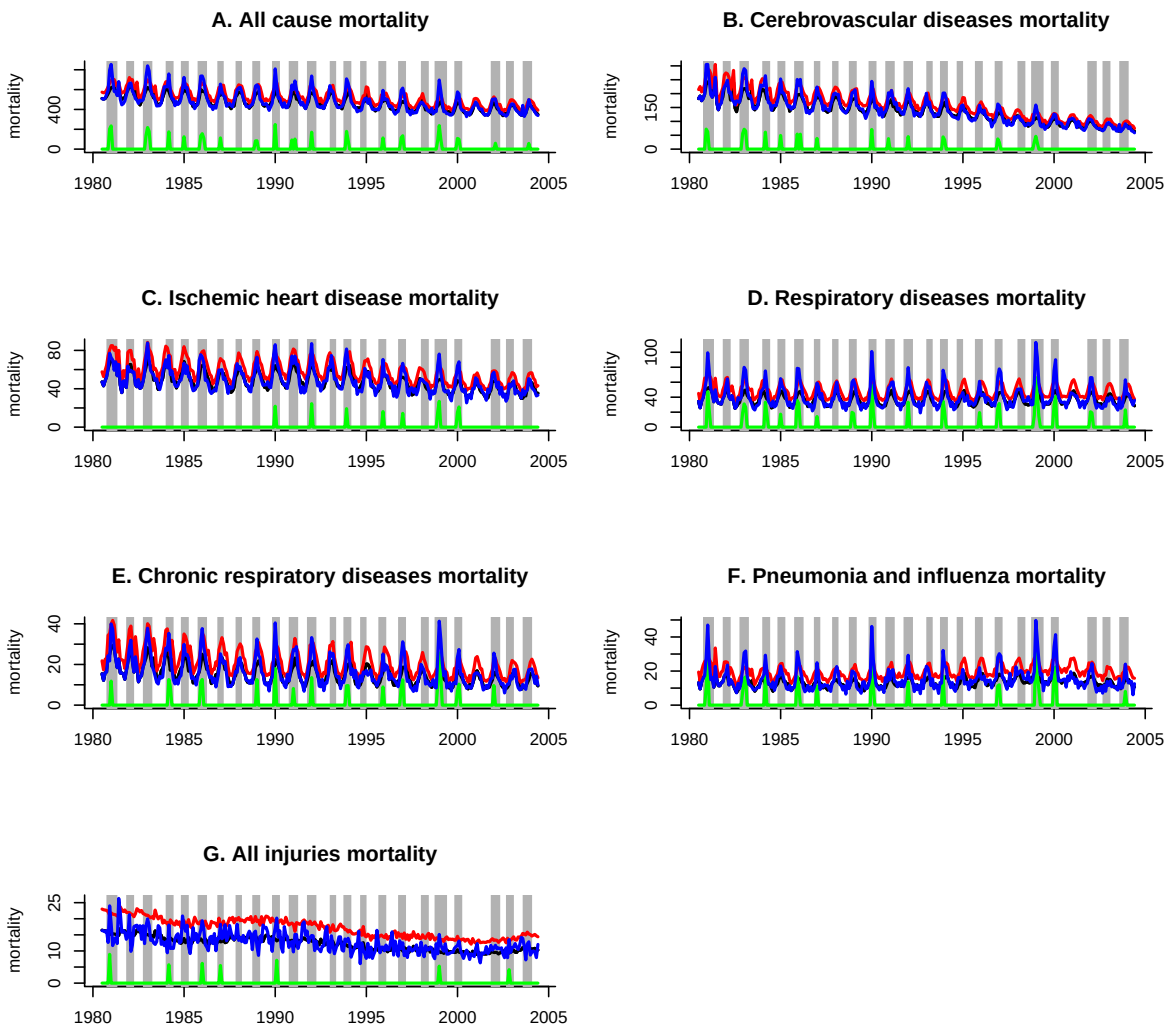

Figure A.6: Mortality rates (blue), mortality baseline (black) and 95% confidence limit (red), estimated excess death rate (green) by month and influenza epidemic periods (grey rectangles) for the study causes of death - age group 75-79

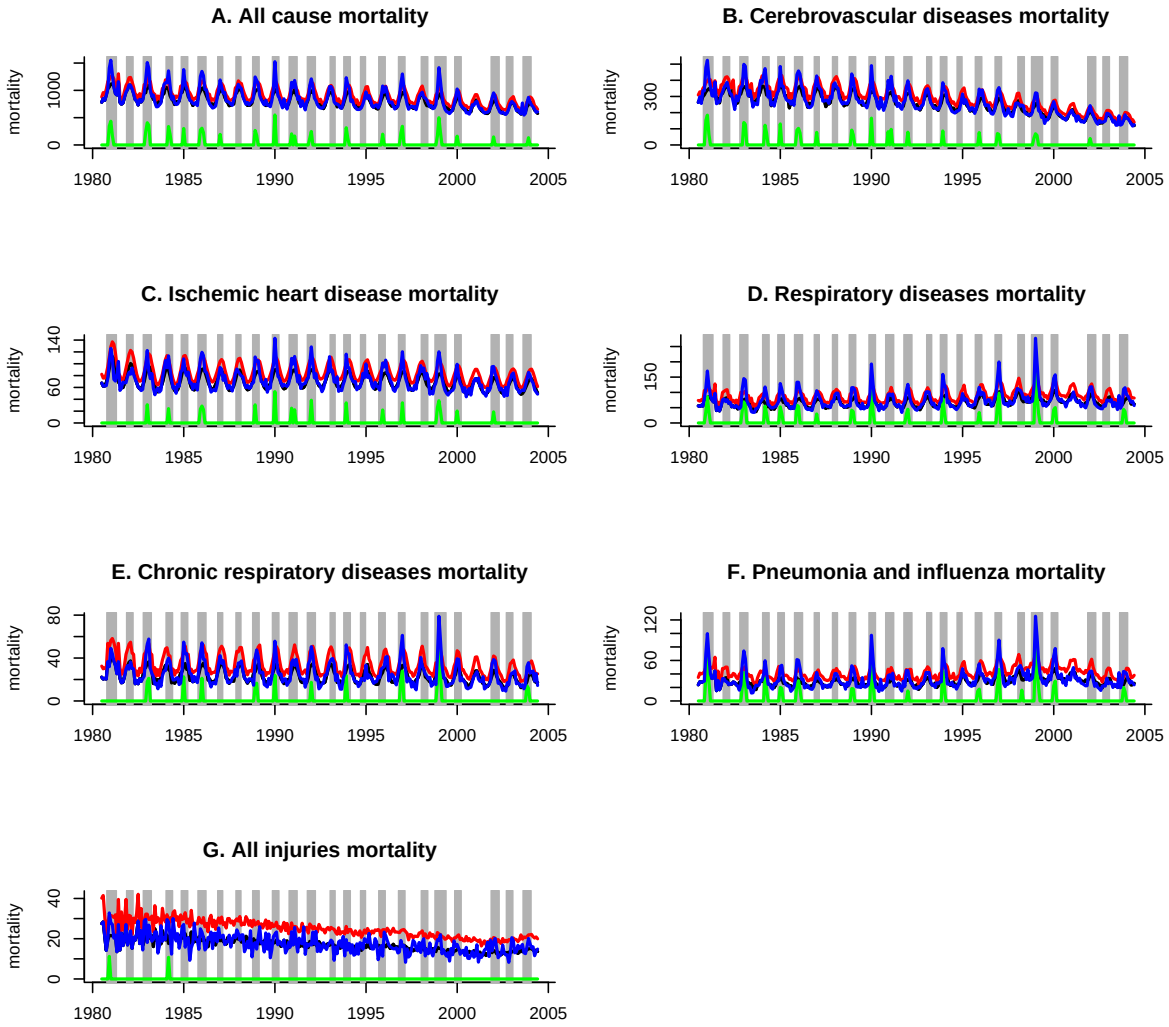

Figure A.7: Mortality rates (blue), mortality baseline (black) and 95% confidence limit (red), estimated excess death rate (green) by month and influenza epidemic periods (grey rectangles) for the study causes of death - age group 80-84

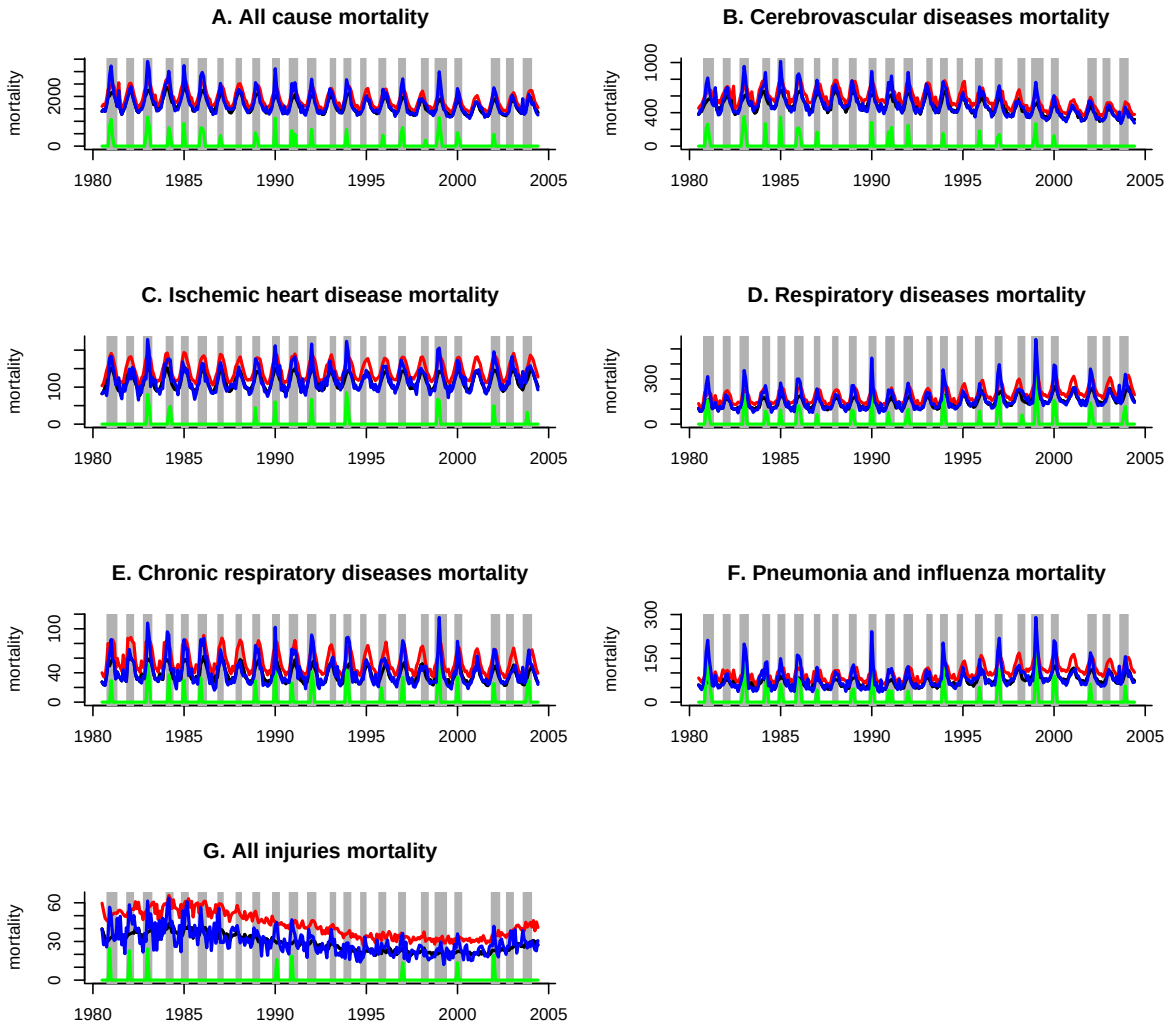

Figure A.8: Mortality rates (blue), mortality baseline (black) and 95% confidence limit (red), estimated excess death rate (green) by month and influenza epidemic periods (grey rectangles) for the study causes of death - age group ≥ 85

Appendix B

Sensitivity analysis

Season	Excess (absolute number)	Age-standardized excess rate 10^5	Proportion of excess (oct-may)	All external	All external rate
1980/1981	47	0.40	1.0%	4547	42.1
1981/1982	30	0.20	0.5%	4783	44.2
1982/1983	15	0.12	0.3%	4650	42.0
1983/1984	29	0.15	0.4%	4628	41.6
1984/1985	0	0.00	0.0%	4472	39.7
1985/1986	16	0.08	0.2%	4265	37.4
1986/1987	15	0.07	0.2%	4461	39.4
1987/1988	0	0.00	0.0%	4420	42.1
1988/1989	15	0.09	0.2%	4393	44.2
1989/1990	32	0.23	0.7%	4097	42.0
1990/1991	30	0.31	0.8%	4255	41.6
1991/1992	0	0.00	0.0%	4338	39.7
1992/1993	0	0.00	0.0%	4059	37.4
1993/1994	0	0.00	0.0%	3672	39.4
1994/1995	0	0.00	0.0%	3654	42.1
1995/1996	30	0.16	0.5%	3655	44.2
1996/1997	18	0.12	0.4%	3590	42.0
1997/1998	18	0.11	0.4%	3509	41.6
1998/1999	37	0.14	0.5%	3318	39.7
1999/2000	27	0.19	0.8%	3199	37.4
2000/2001	0	0.00	0.0%	3055	39.4
2001/2002	277	2.56	8.4%	3865	42.1
2002/2003	26	0.17	0.6%	3684	44.2
2003/2004	0	0.00	0.0%	3518	42.0
Mean	28	0.21	0.66%	4004	41.2

Table B.1: Sensitivity analysis: excess deaths and age-standardized excess death rates from injuries that are "attributable to influenza" by the used method. Injuries comprise all external causes of death.

Appendix C

Hidden Markov Models parameters convergence

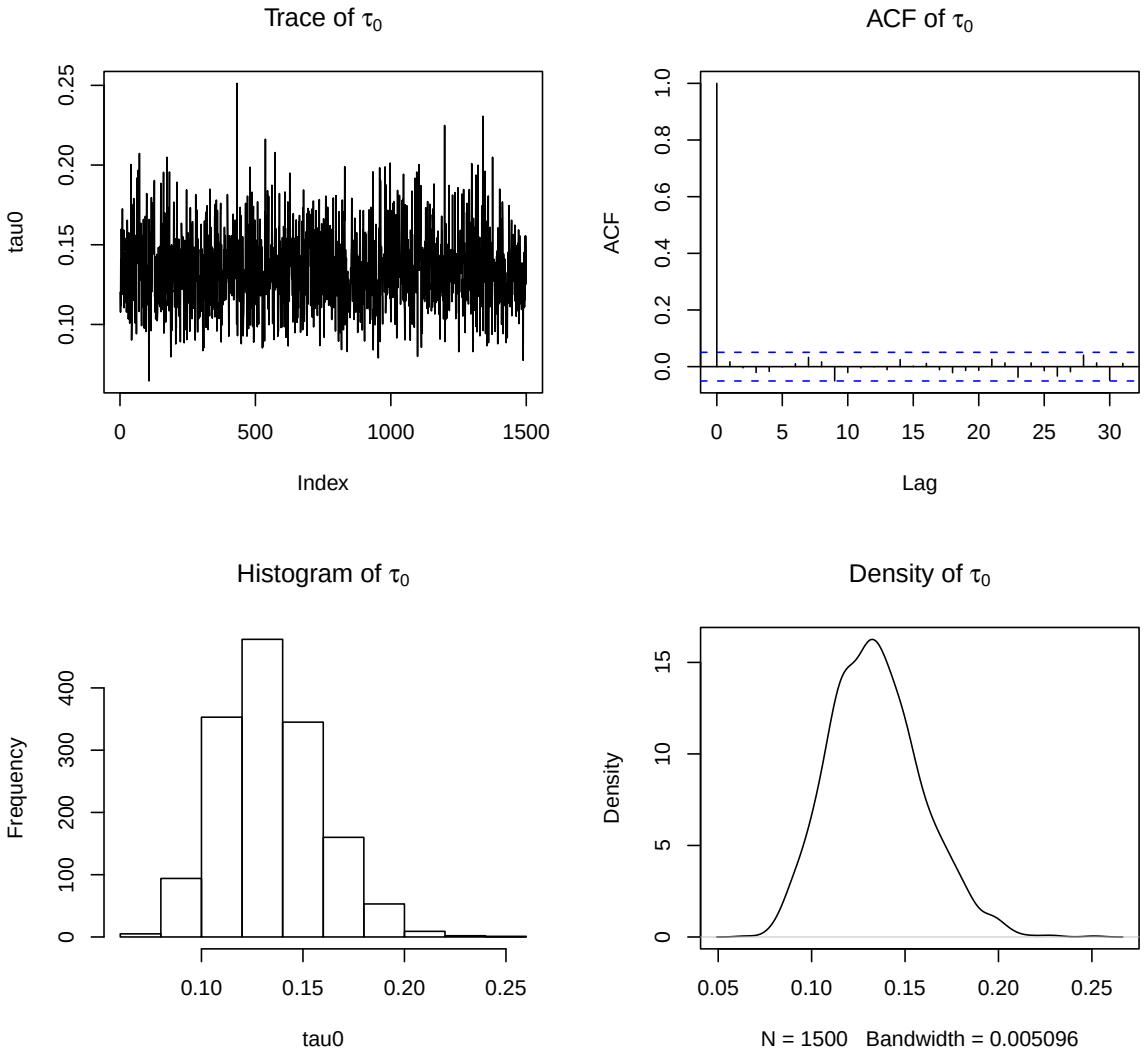

Figure C.1: Trace, autocorrelation function, histogram and density of τ_0 parameter of model 0

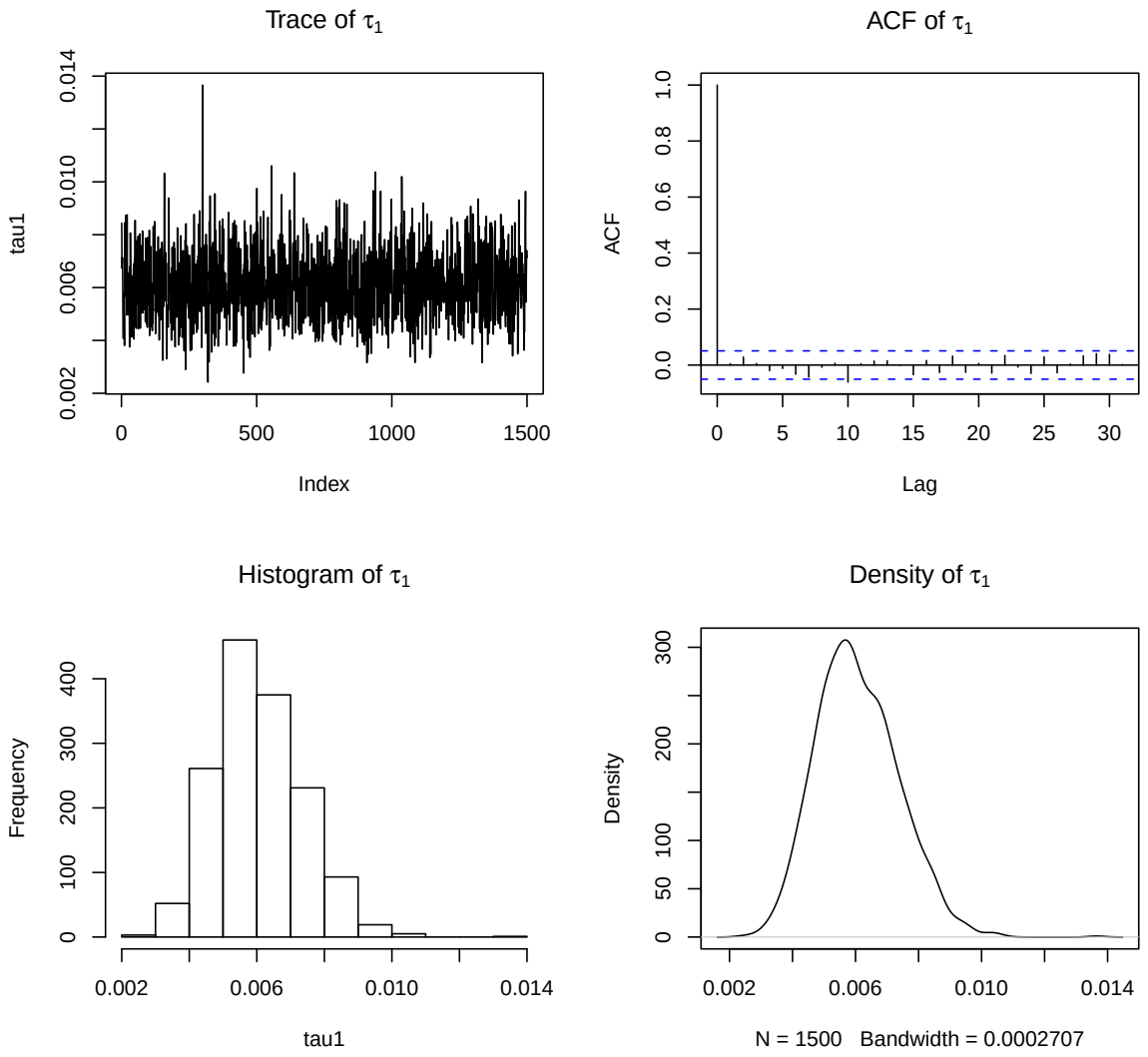

Figure C.2: Trace, autocorrelation function, histogram and density of τ_1 parameter of model 0

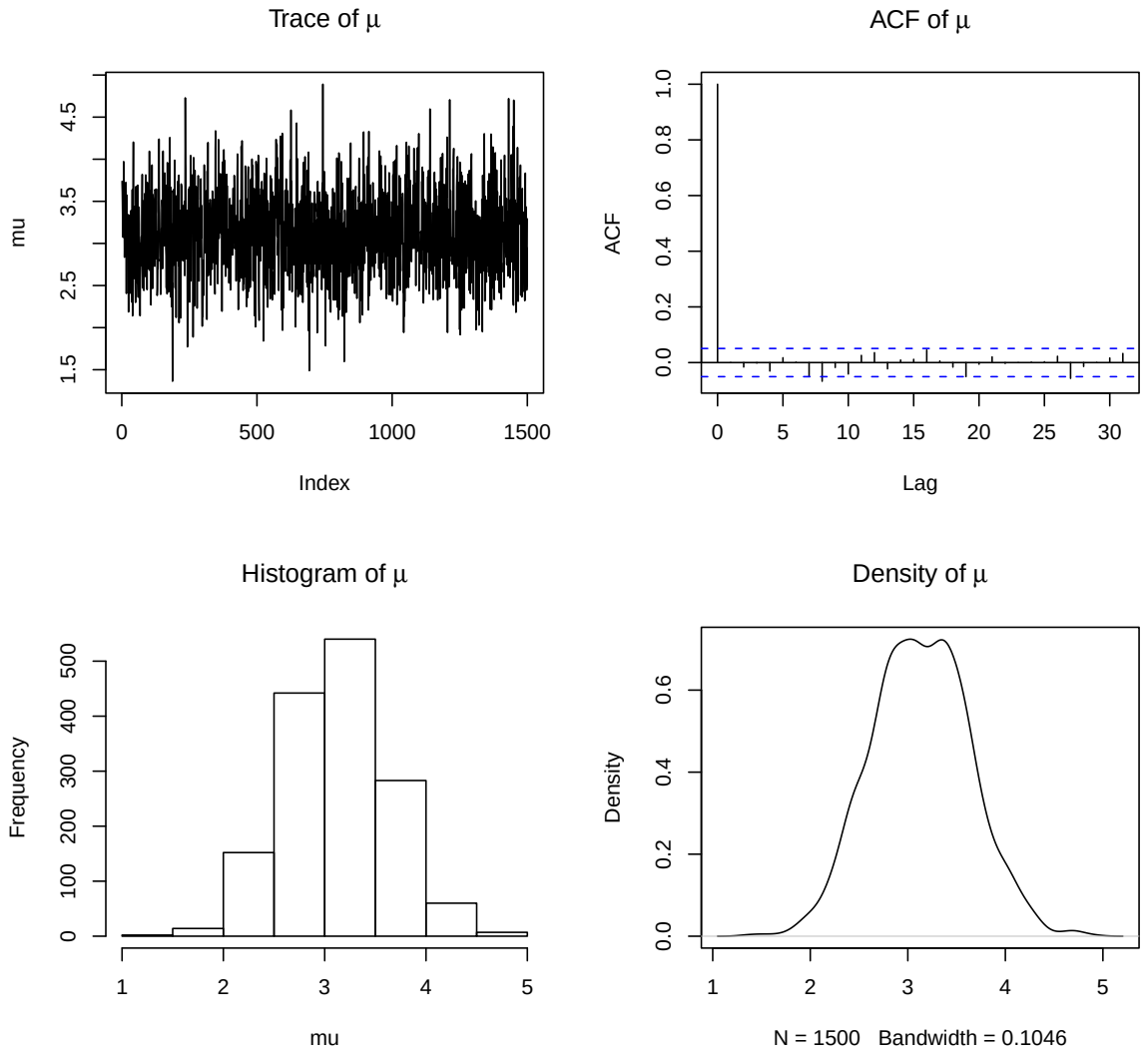

Figure C.3: Trace, autocorrelation function, histogram and density of μ parameter of model 0

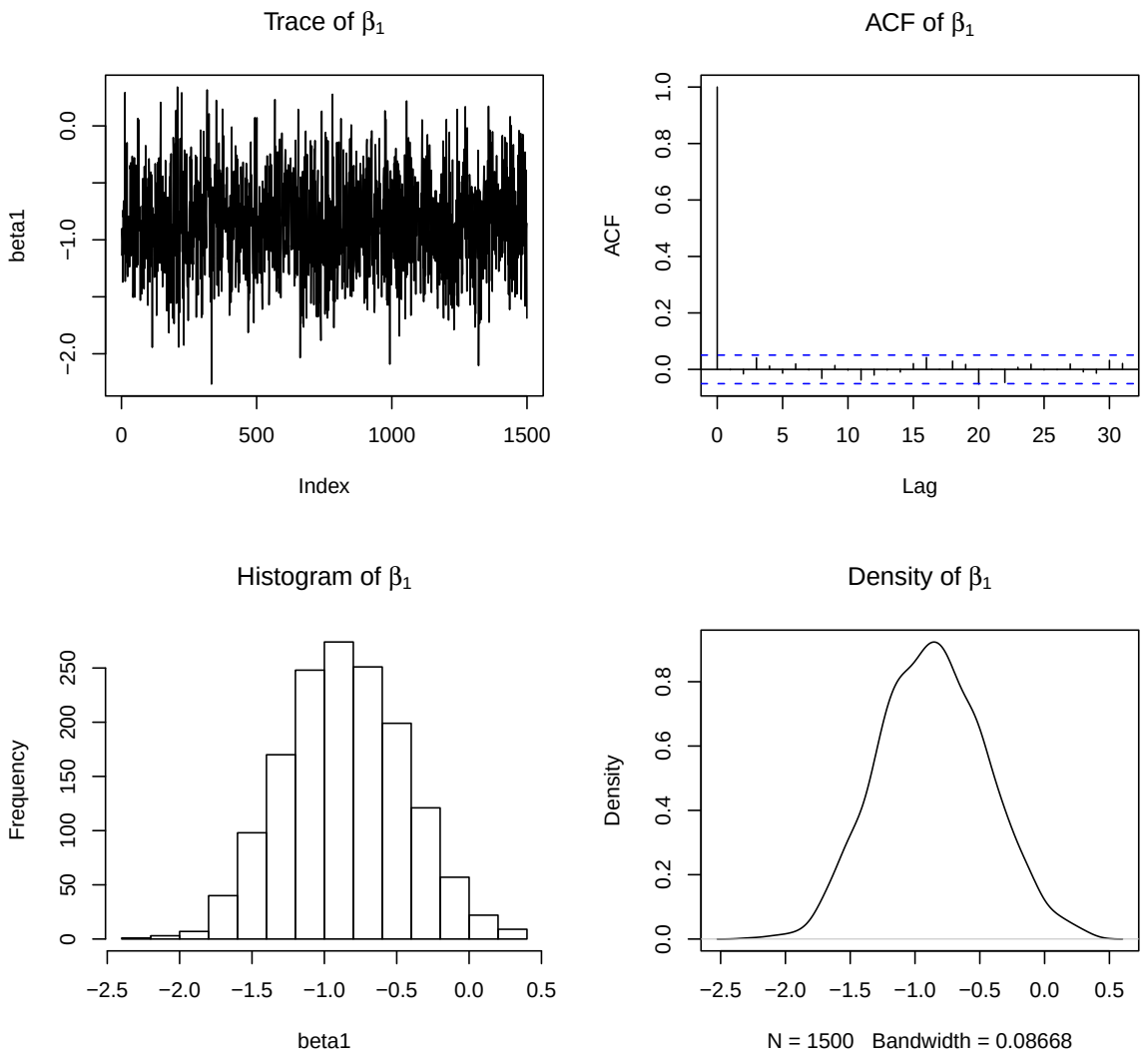

Figure C.4: Trace, autocorrelation function, histogram and density of β_1 parameter of model 0

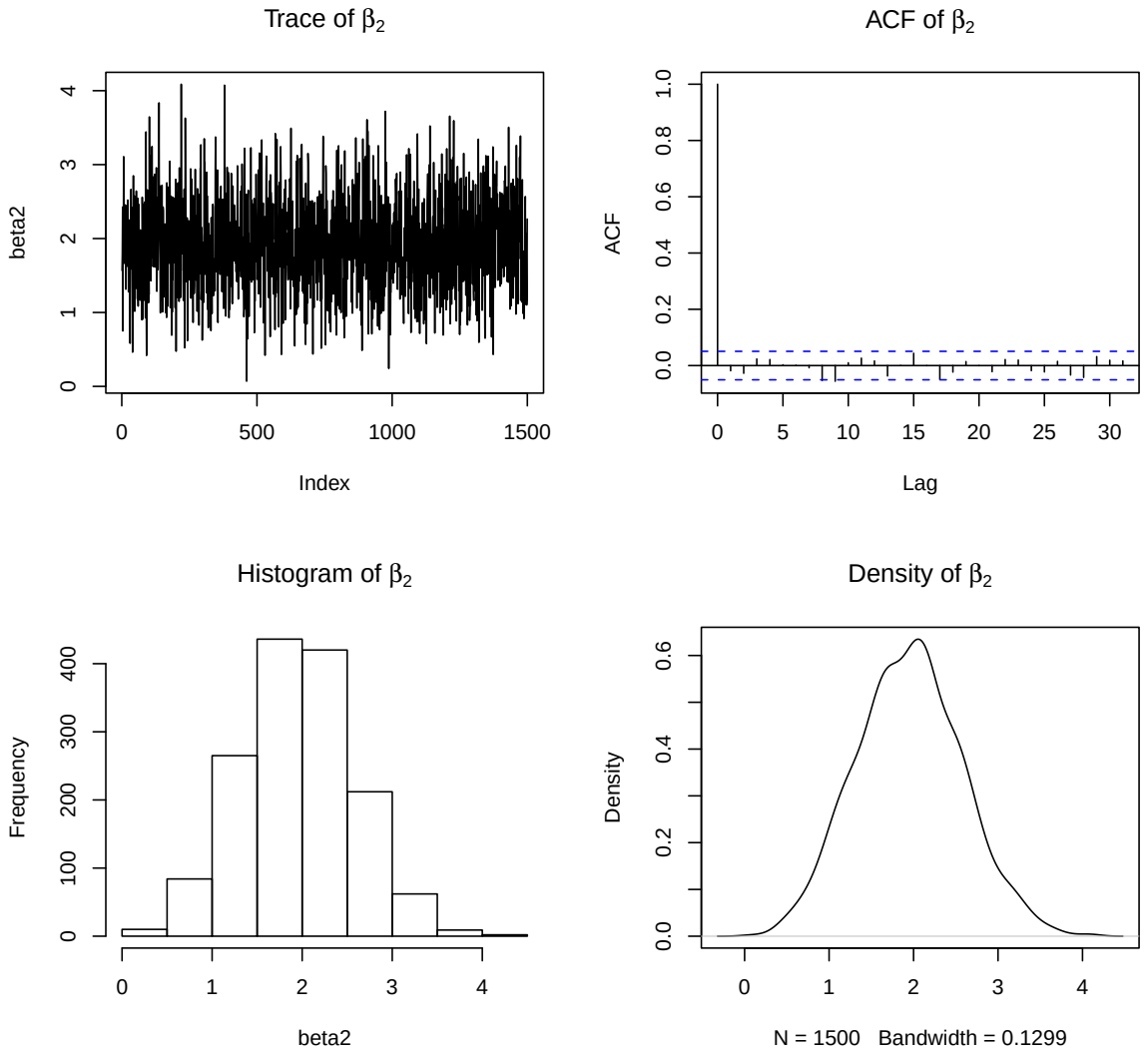

Figure C.5: Trace, autocorrelation function, histogram and density of β_2 parameter of model 0

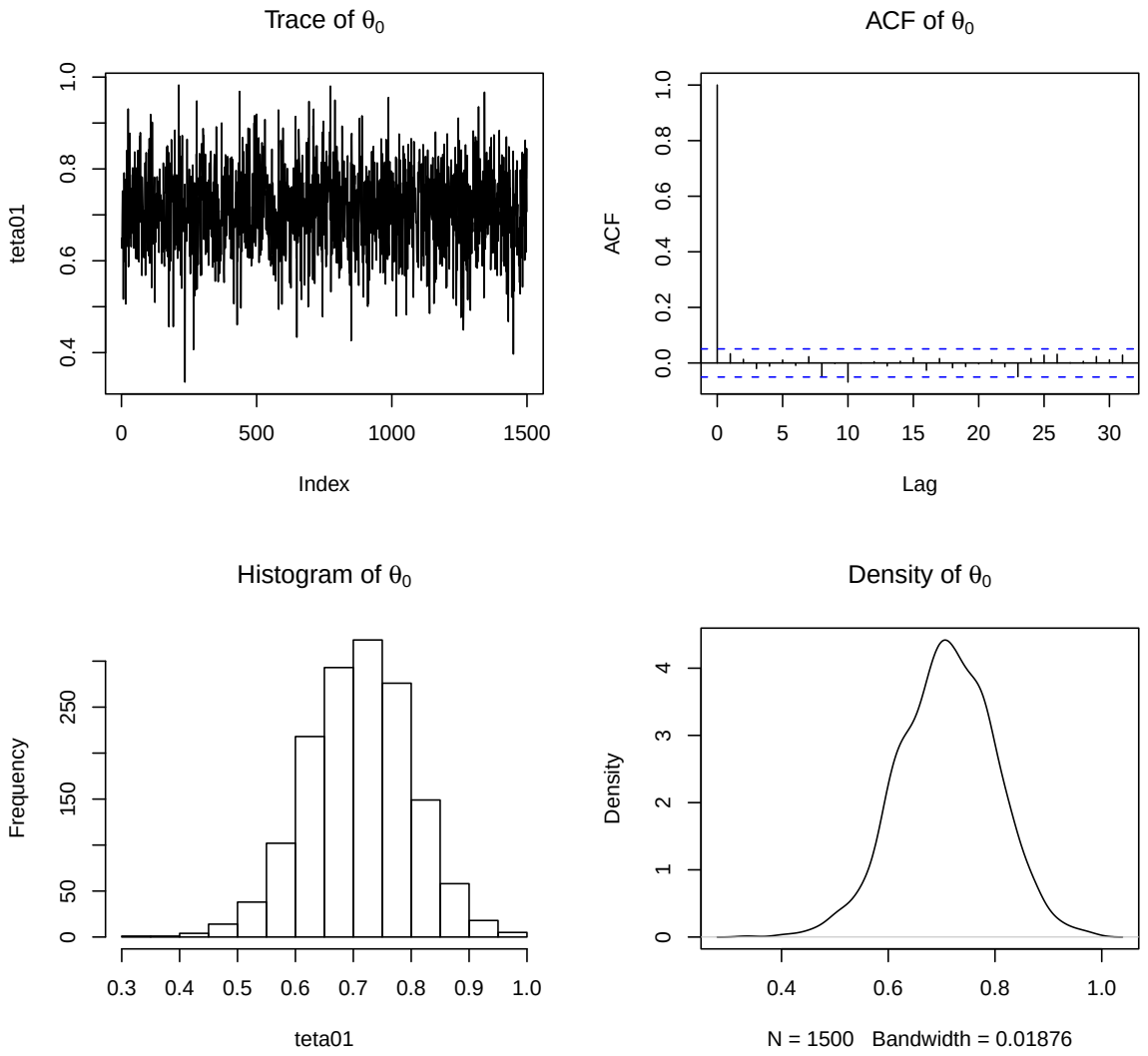

Figure C.6: Trace, autocorrelation function, histogram and density of θ_0 parameter of model 0

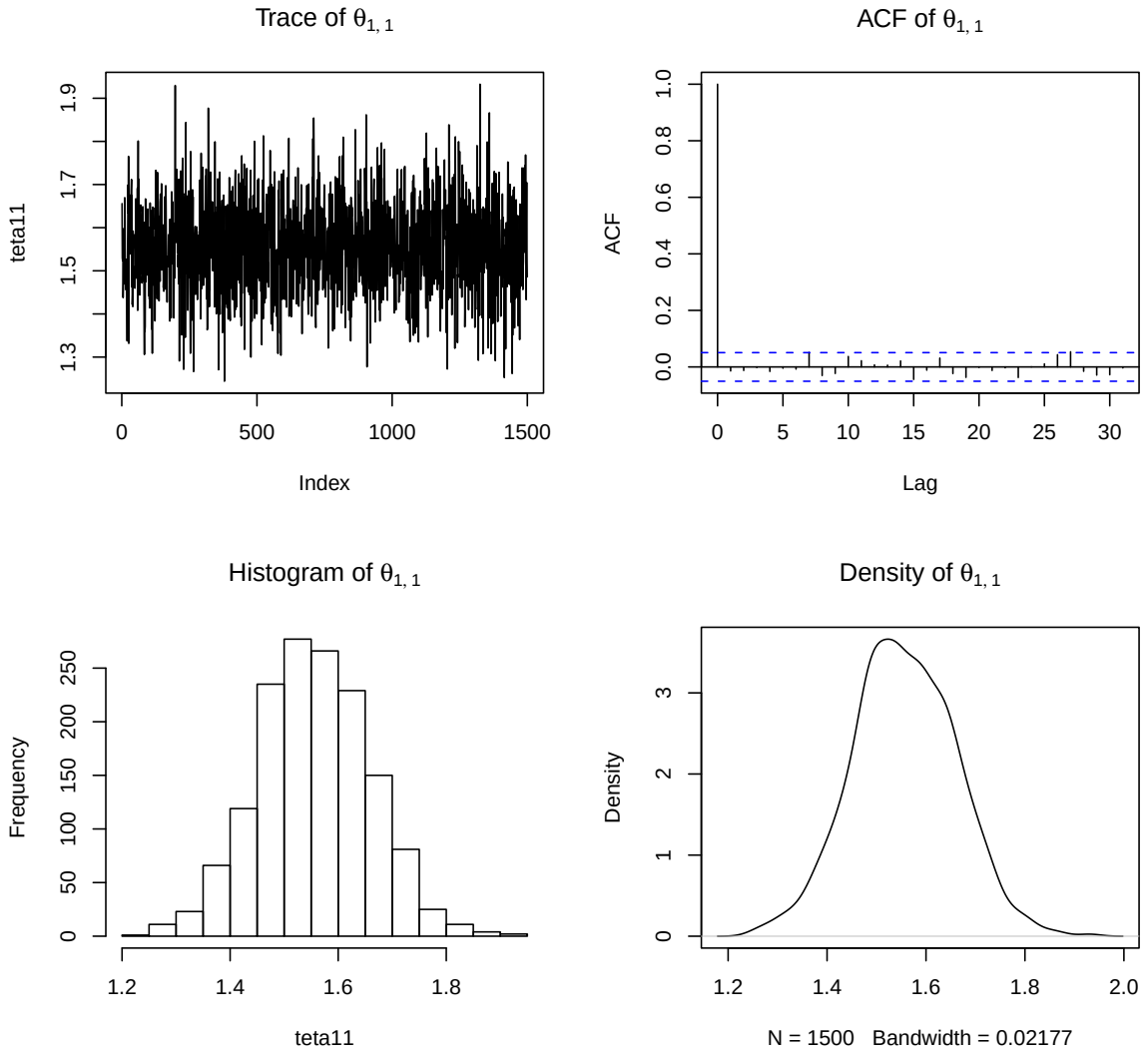

Figure C.7: Trace, autocorrelation function, histogram and density of $\theta_{1,1}$ parameter of model 0

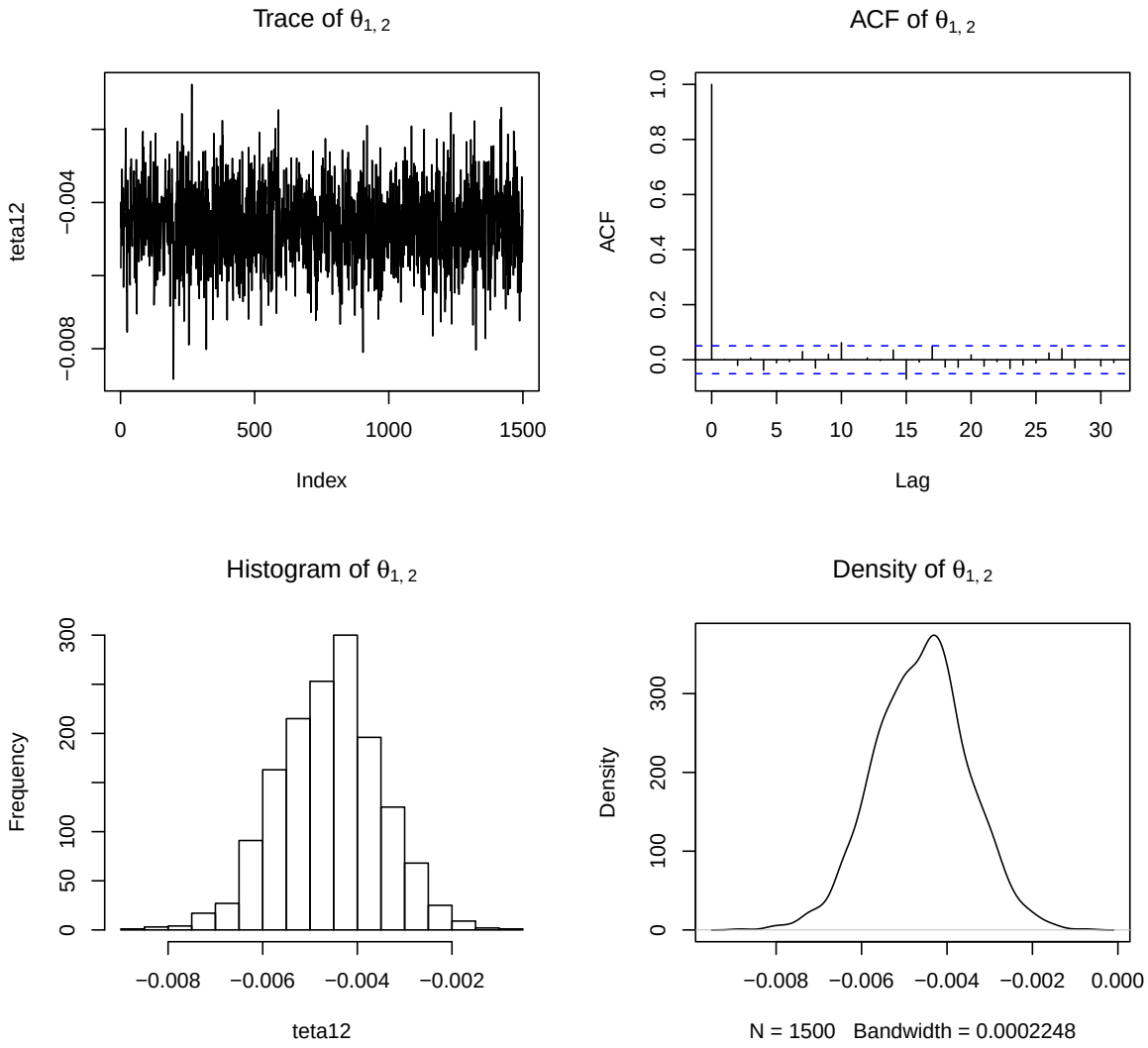

Figure C.8: Trace, autocorrelation function, histogram and density of $\theta_{1,2}$ parameter of model 0

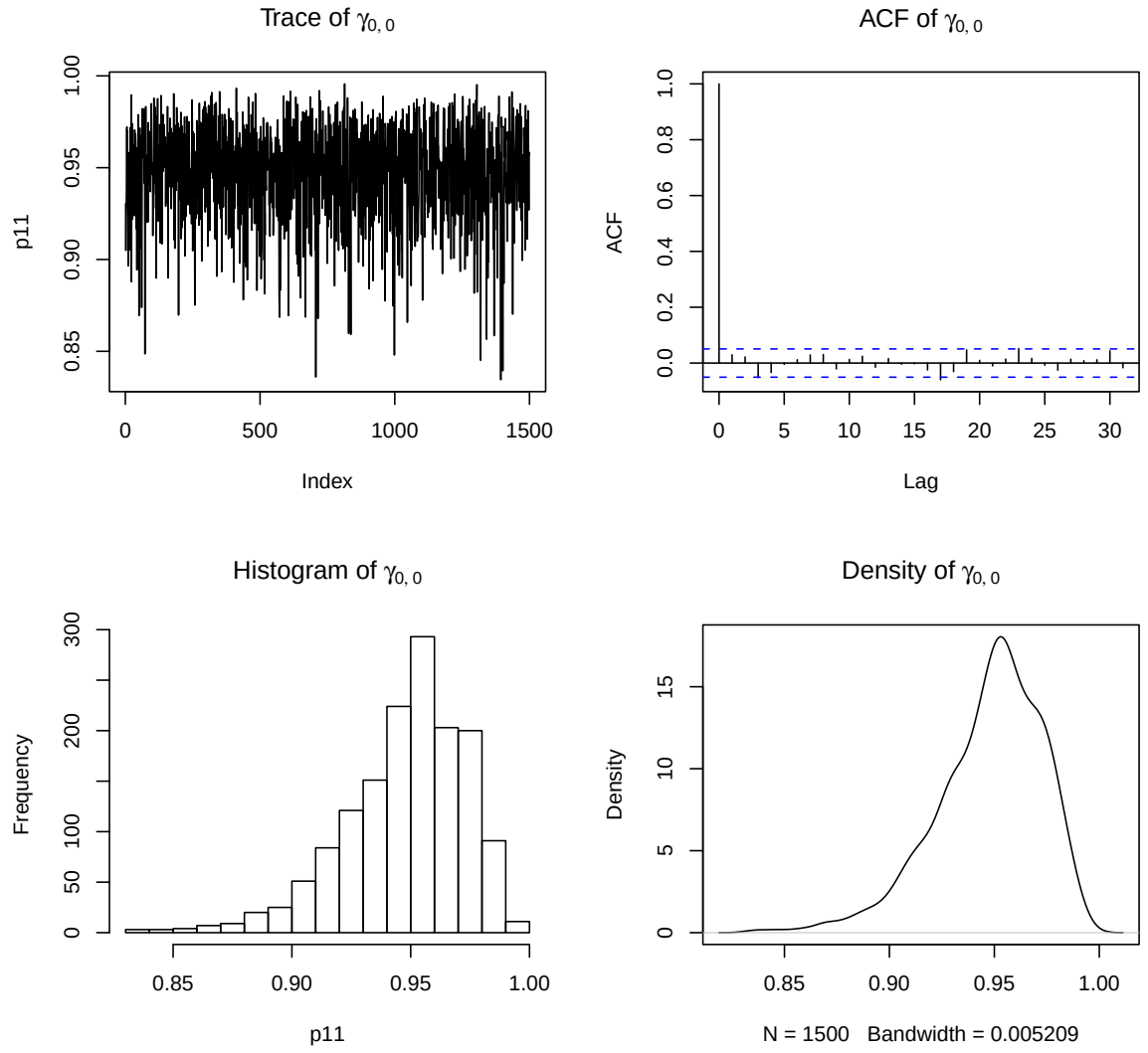

Figure C.9: Trace, autocorrelation function, histogram and density of $\gamma_{0,0}$ parameter of model 0

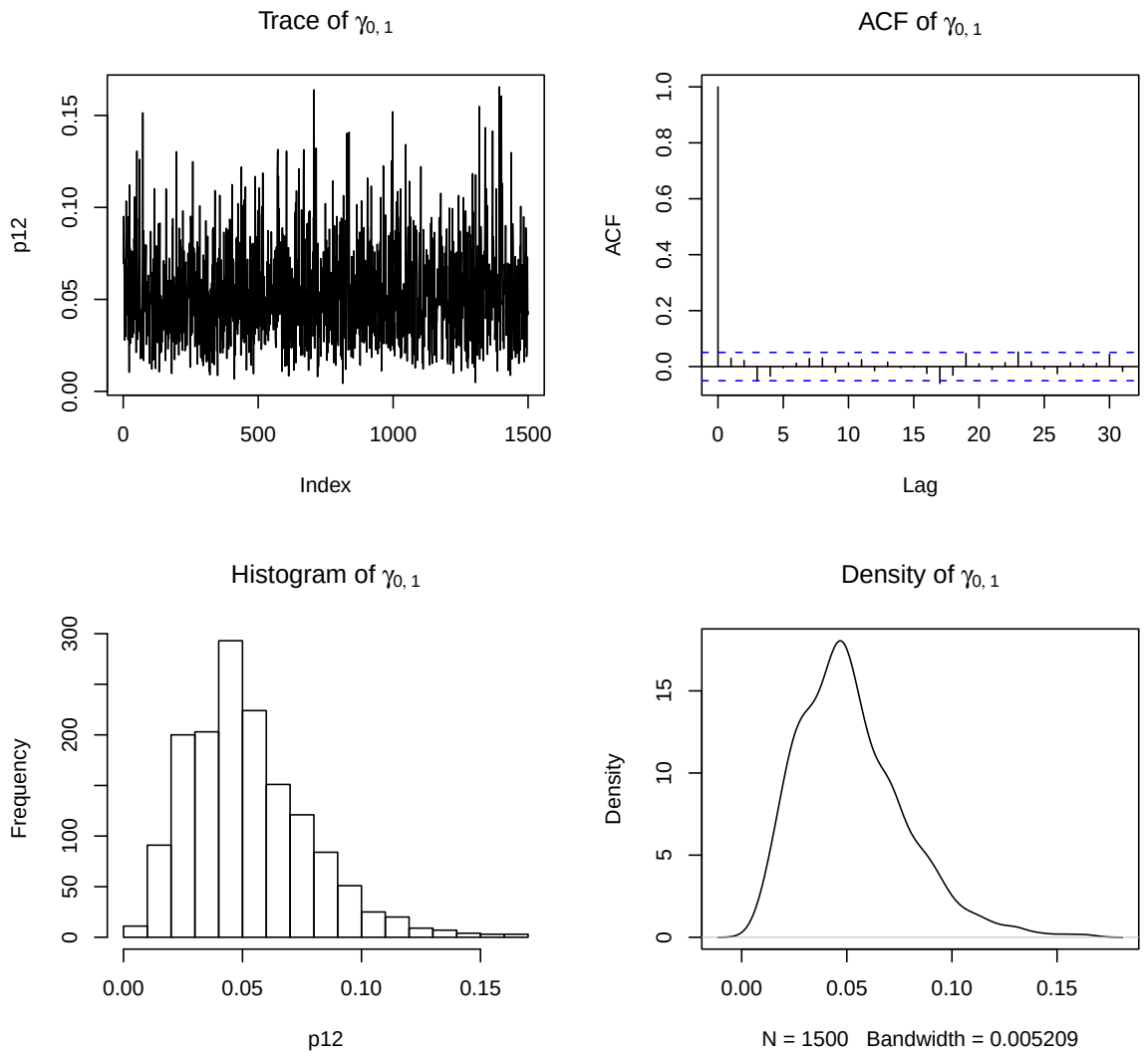

Figure C.10: Trace, autocorrelation function, histogram and density of $\gamma_{0,1}$ parameter of model 0

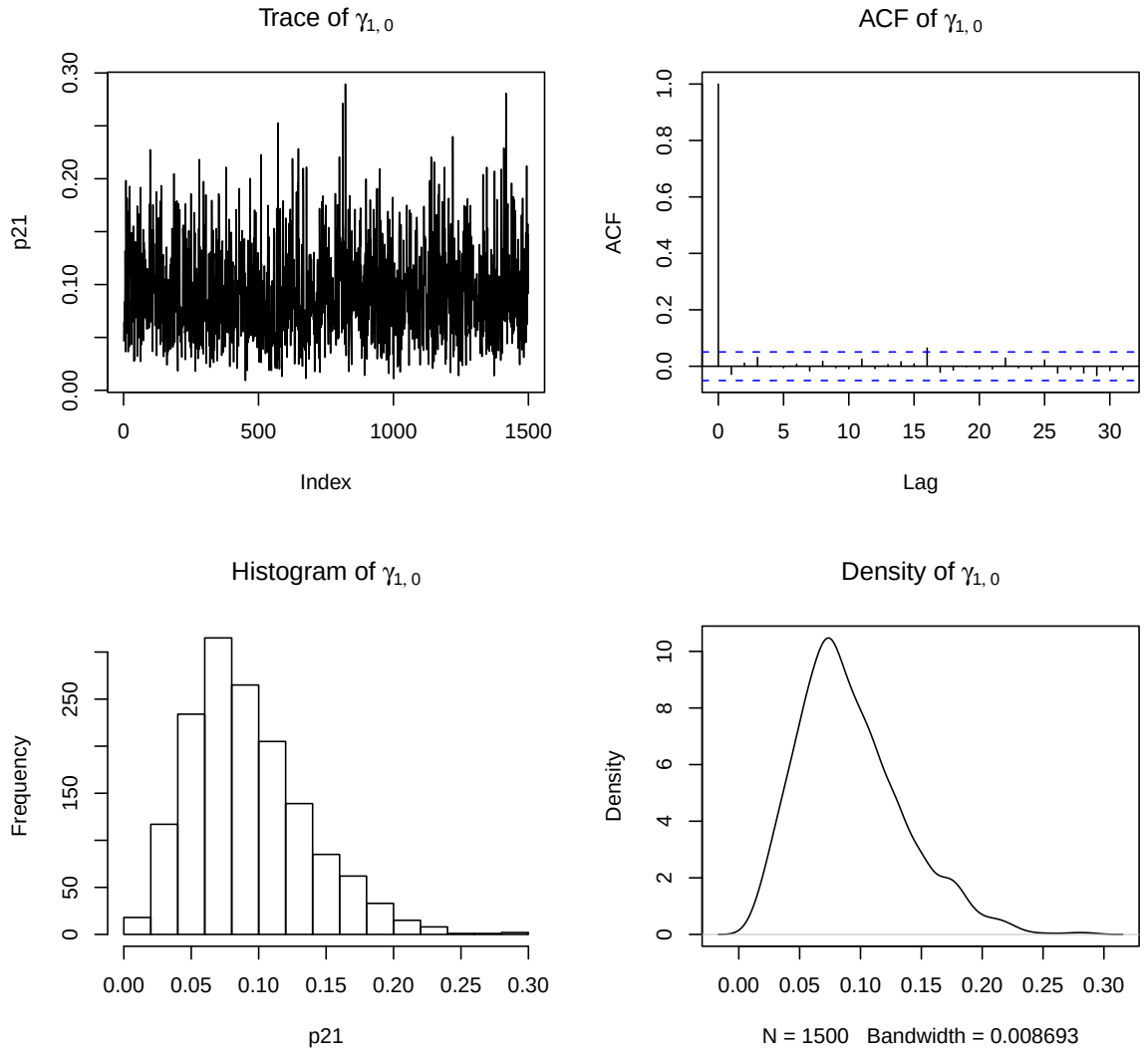

Figure C.11: Trace, autocorrelation function, histogram and density of $\gamma_{1,0}$ parameter of model 0

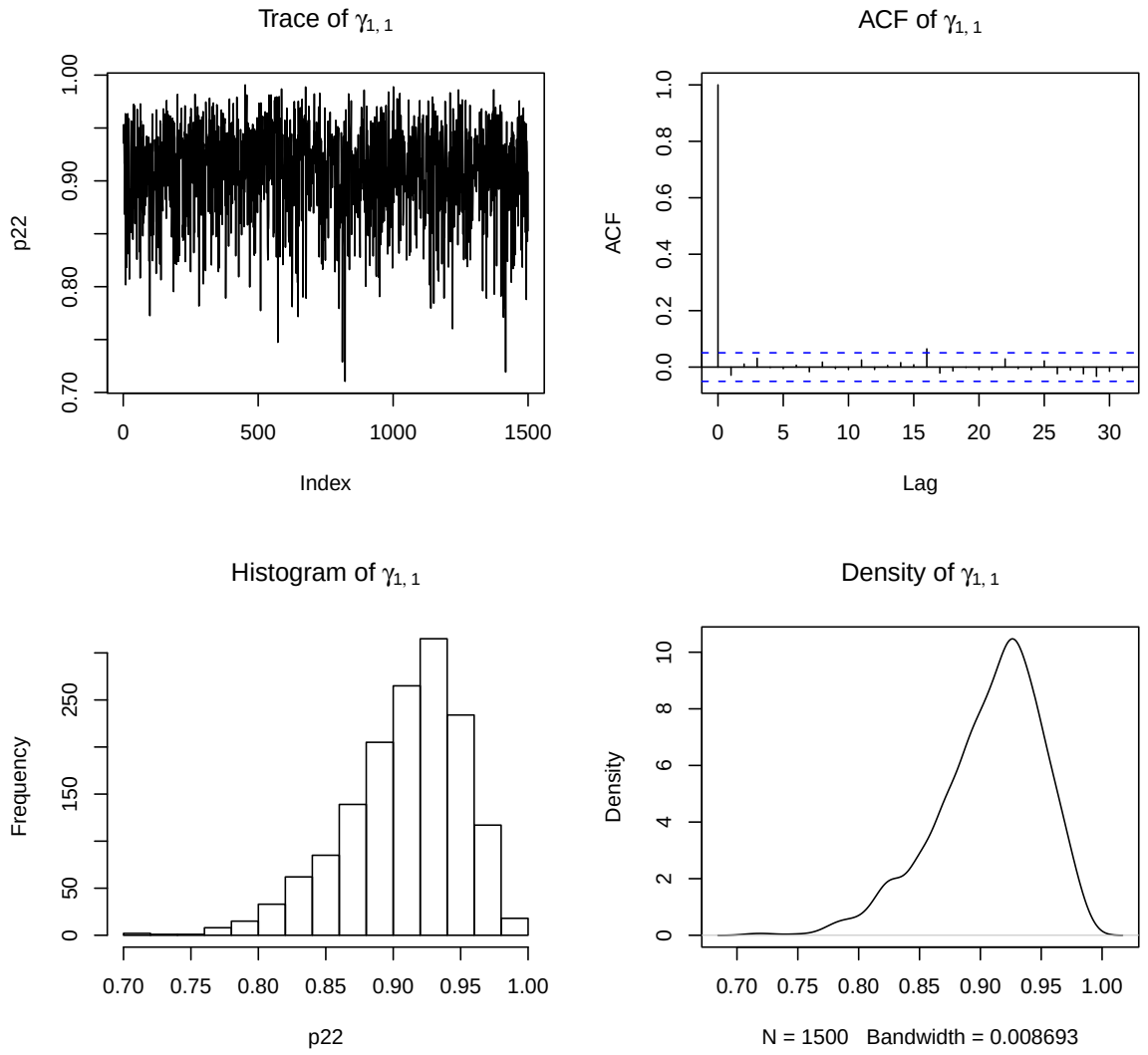

Figure C.12: Trace, autocorrelation function, histogram and density of $\gamma_{1,1}$ parameter of model 0

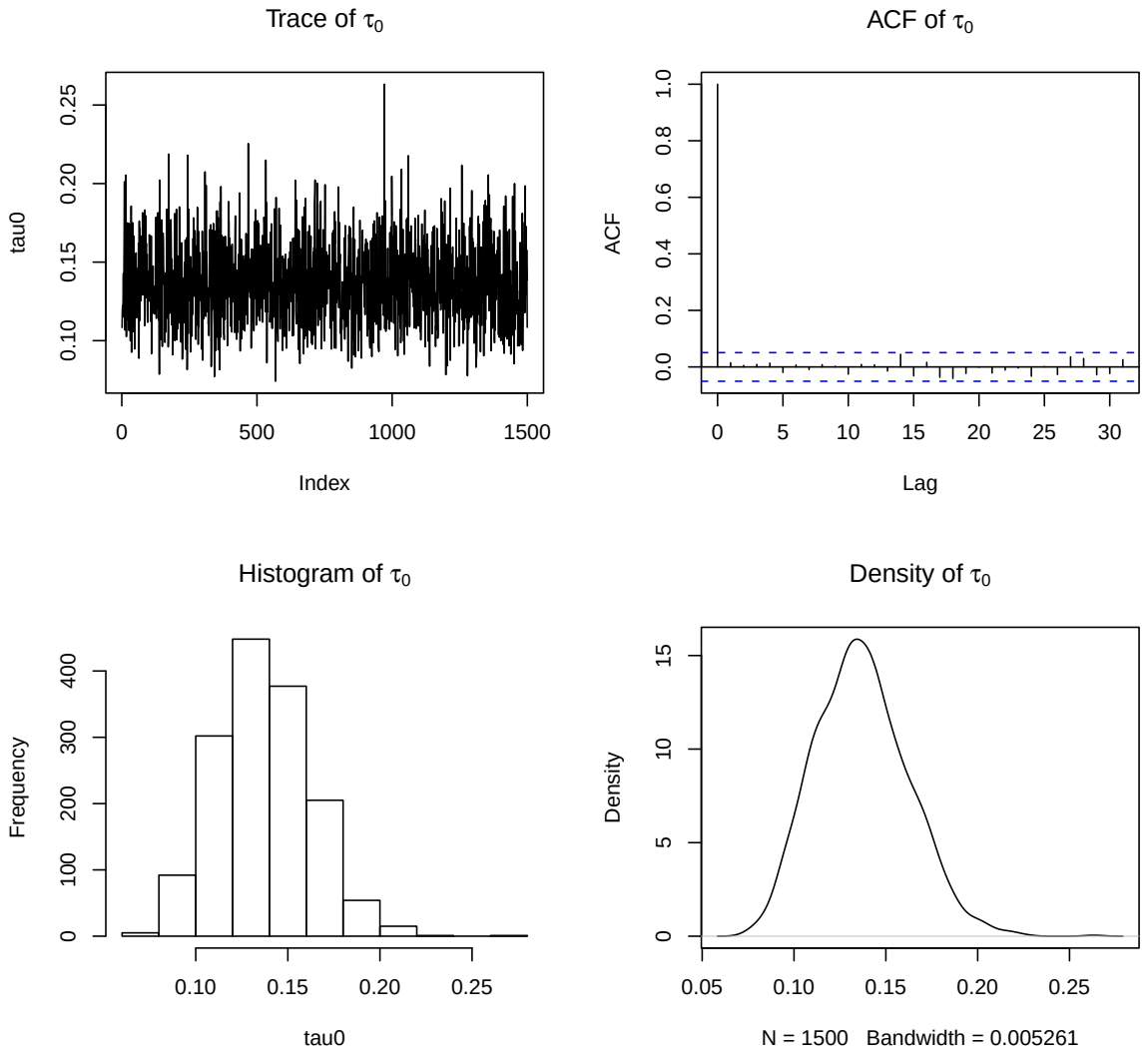

Figure C.13: Trace, autocorrelation function, histogram and density of τ_0 parameter of model 1

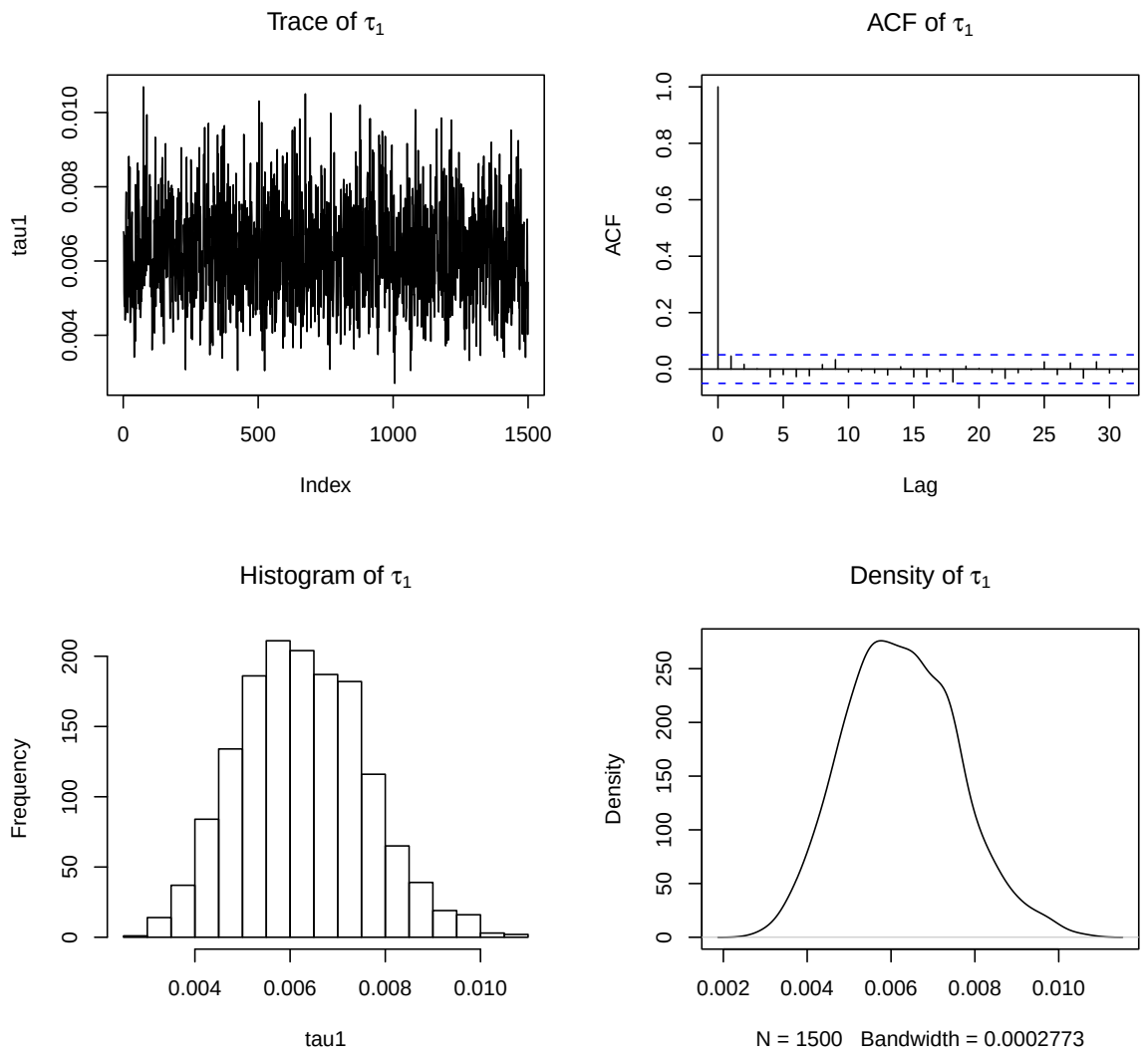

Figure C.14: Trace, autocorrelation function, histogram and density of τ_1 parameter of model 1

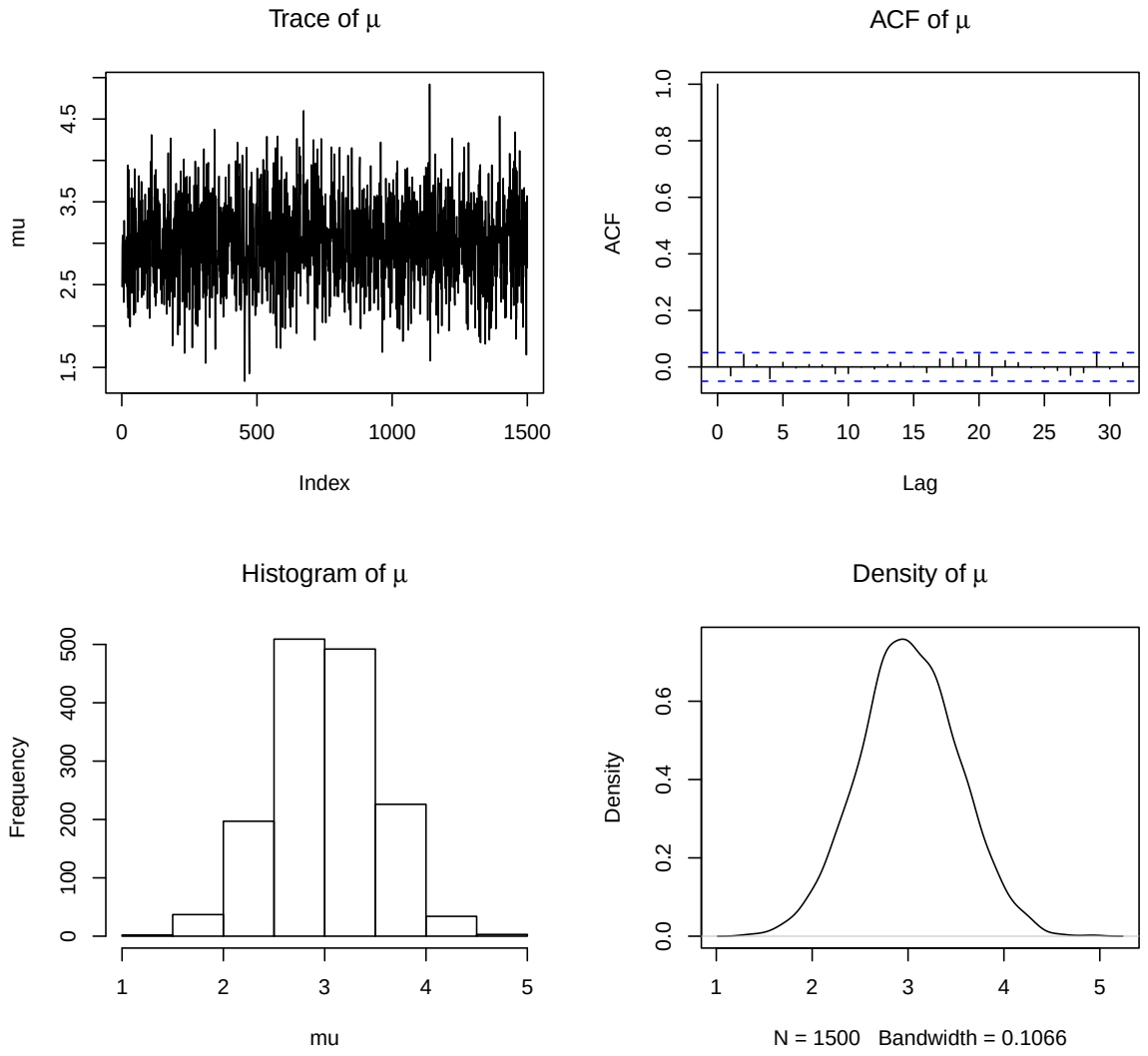

Figure C.15: Trace, autocorrelation function, histogram and density of μ parameter of model 1

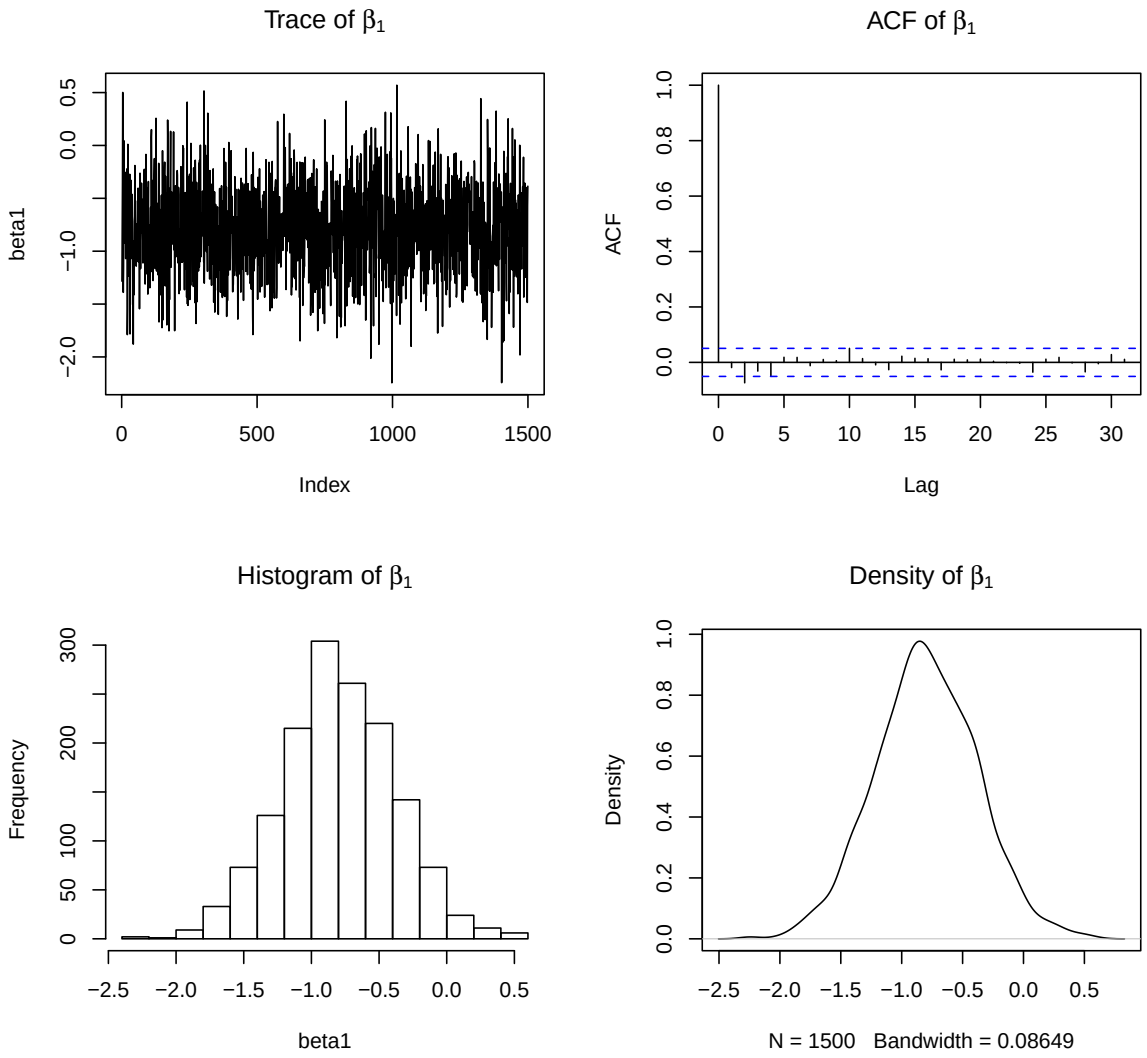

Figure C.16: Trace, autocorrelation function, histogram and density of β_1 parameter of model 1

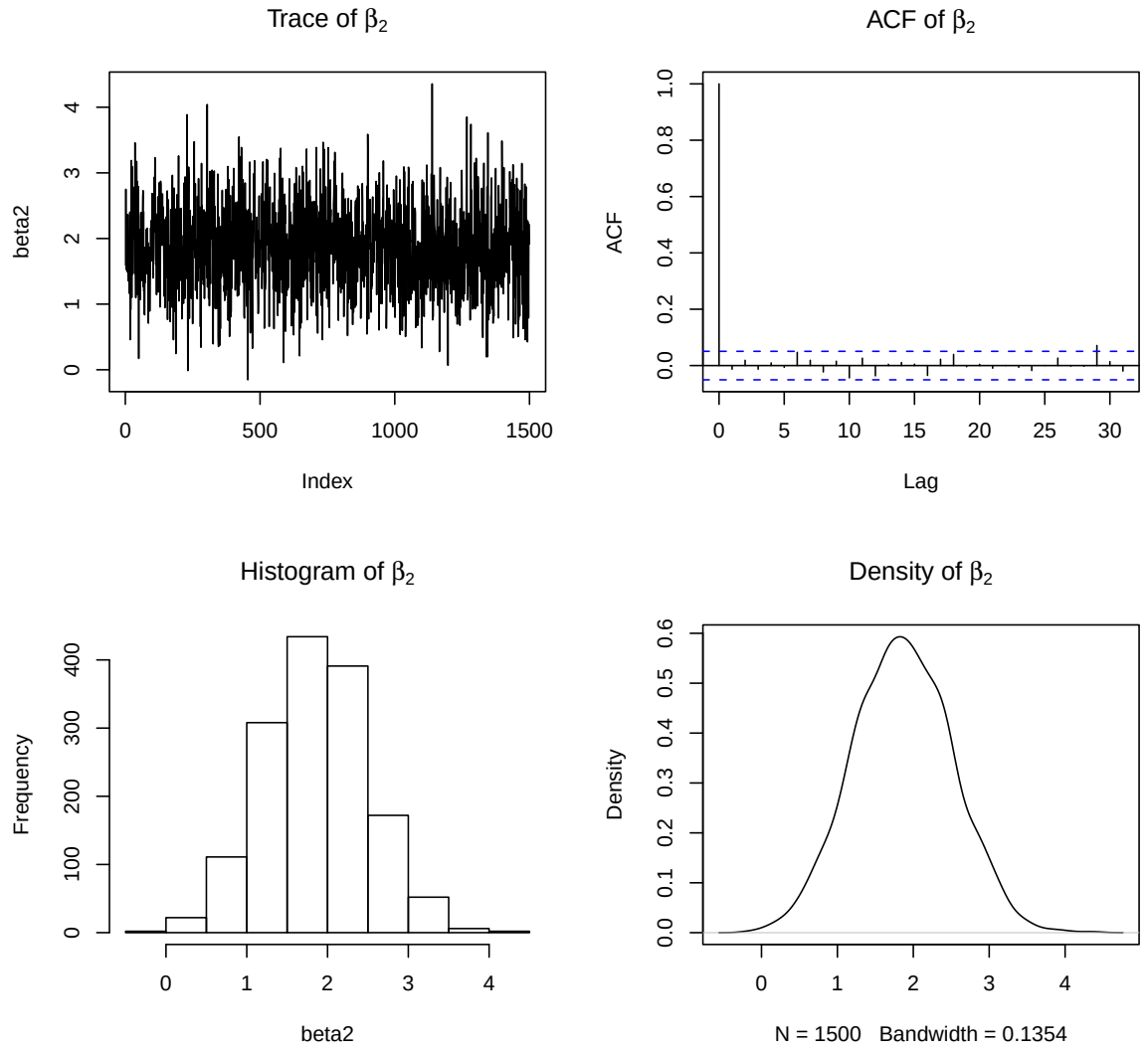

Figure C.17: Trace, autocorrelation function, histogram and density of β_2 parameter of model 1

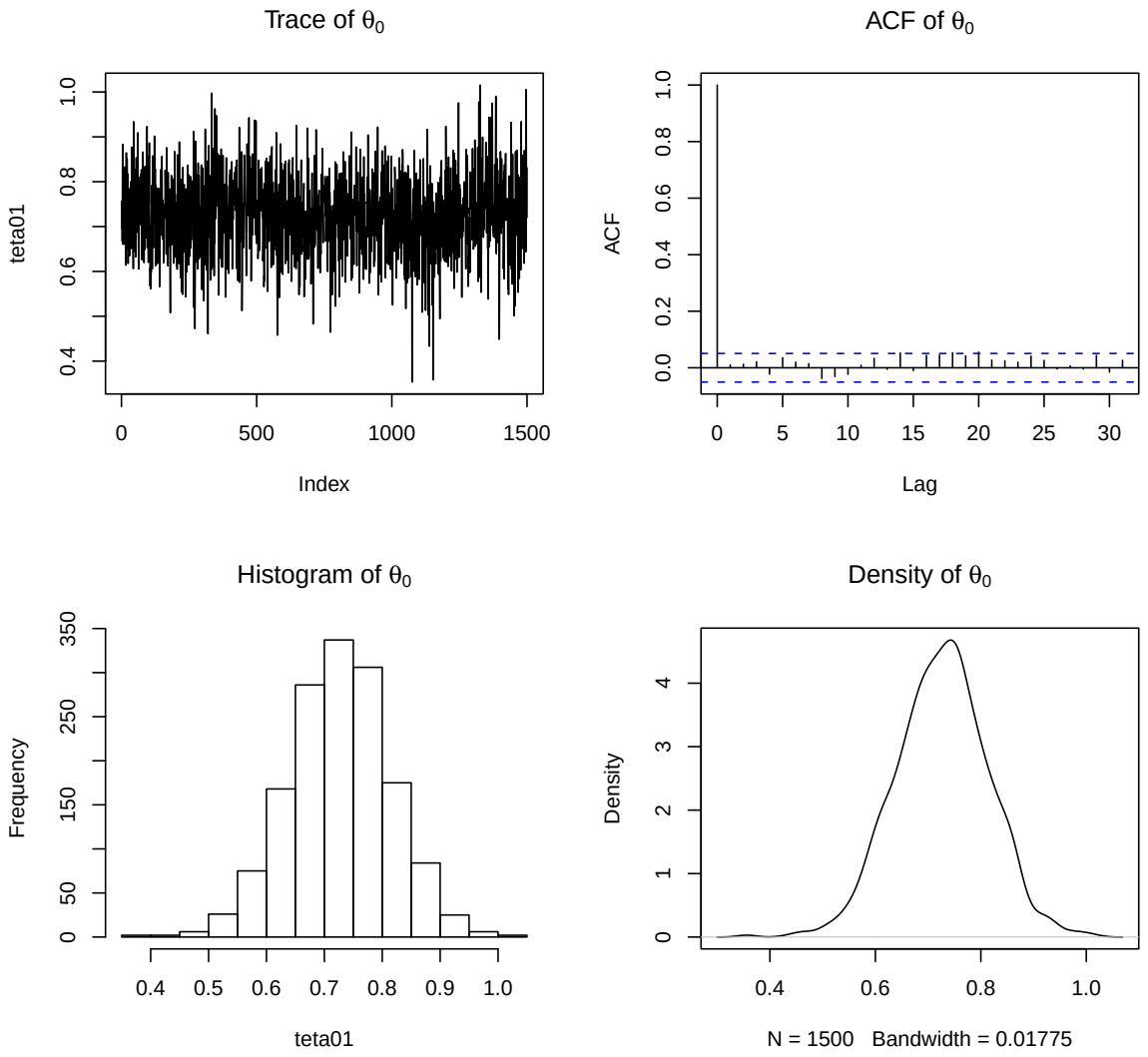

Figure C.18: Trace, autocorrelation function, histogram and density of θ_0 parameter of model 1

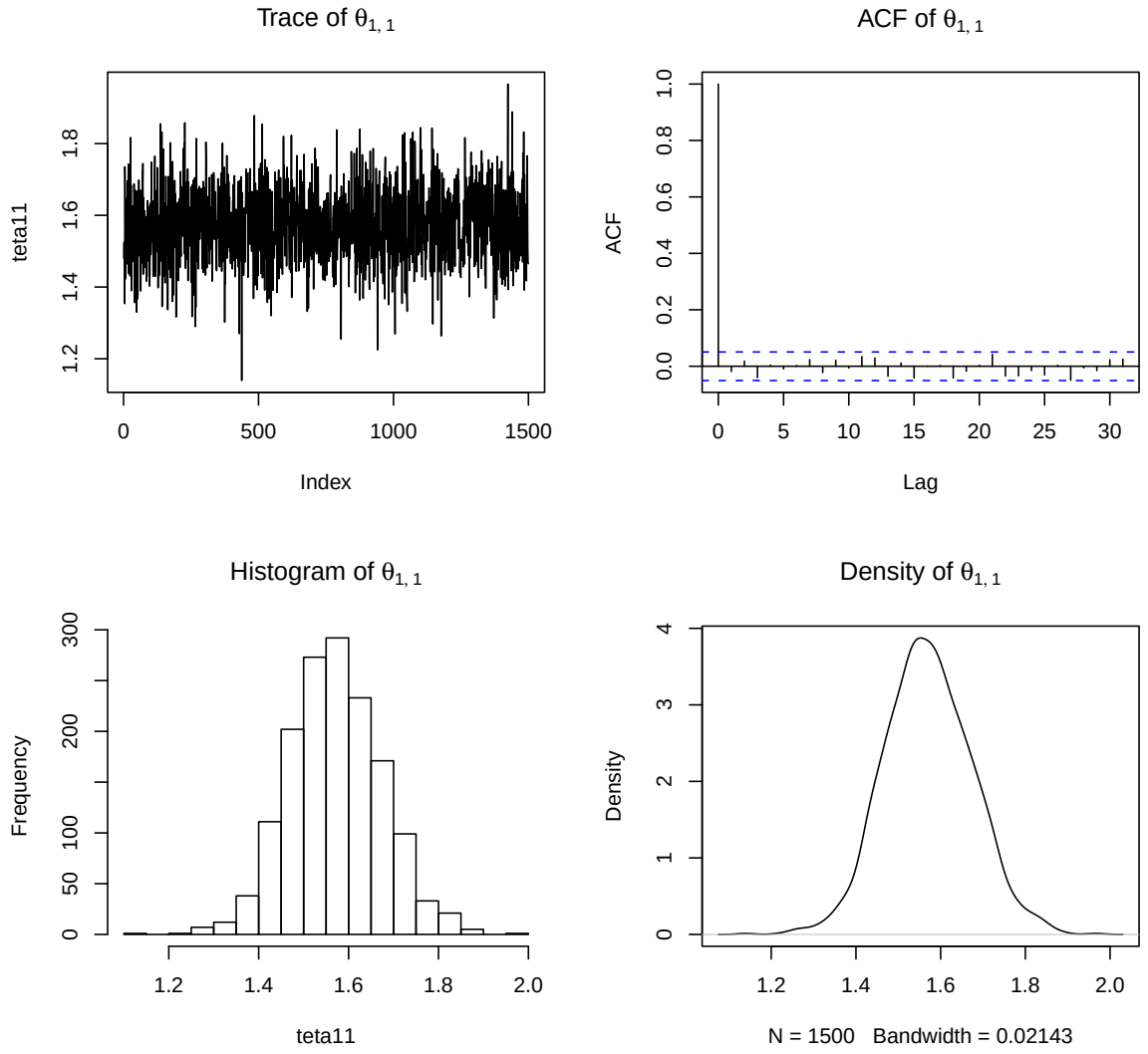

Figure C.19: Trace, autocorrelation function, histogram and density of $\theta_{1,1}$ parameter of model 1

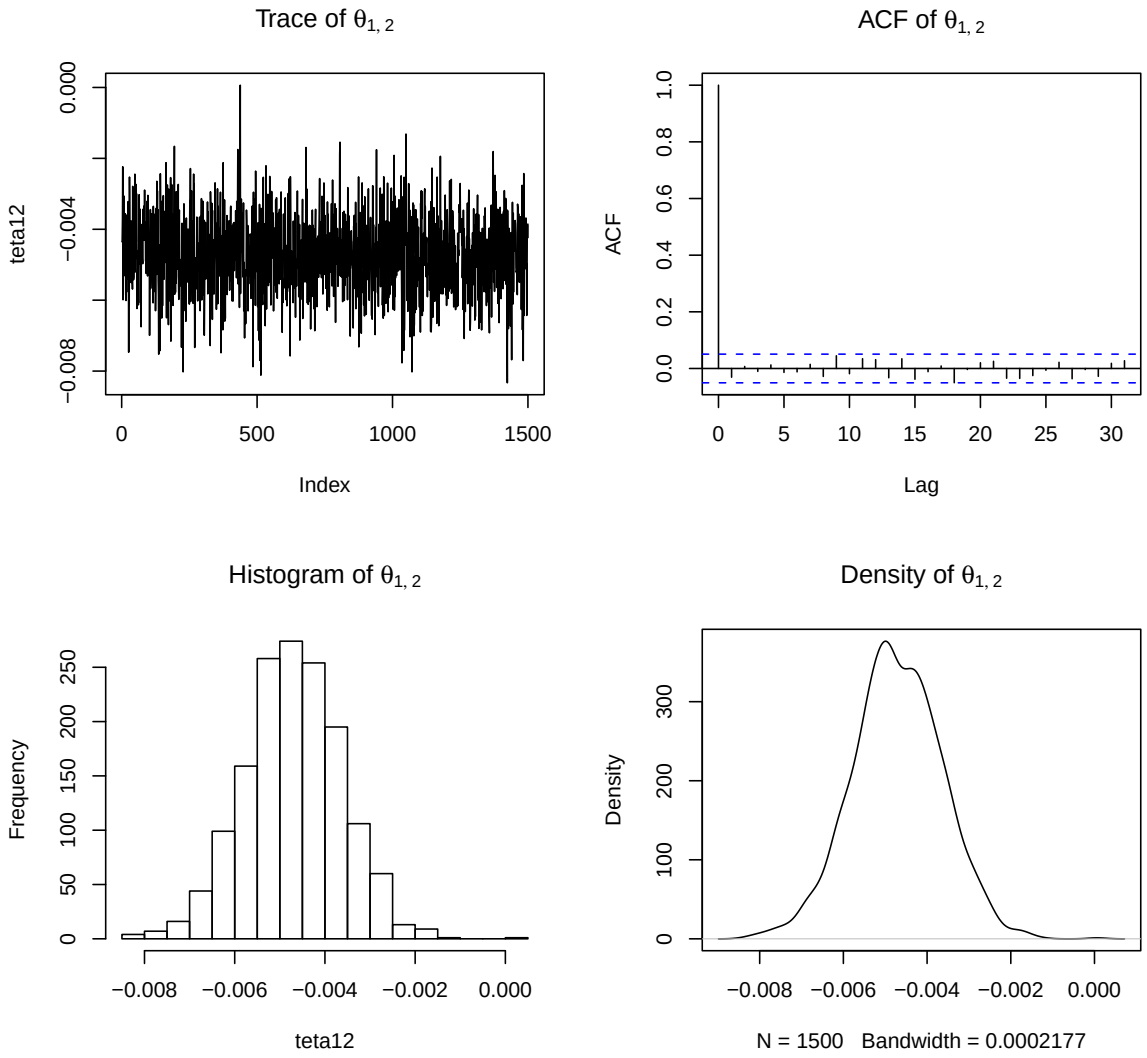

Figure C.20: Trace, autocorrelation function, histogram and density of $\theta_{1,2}$ parameter of model 1

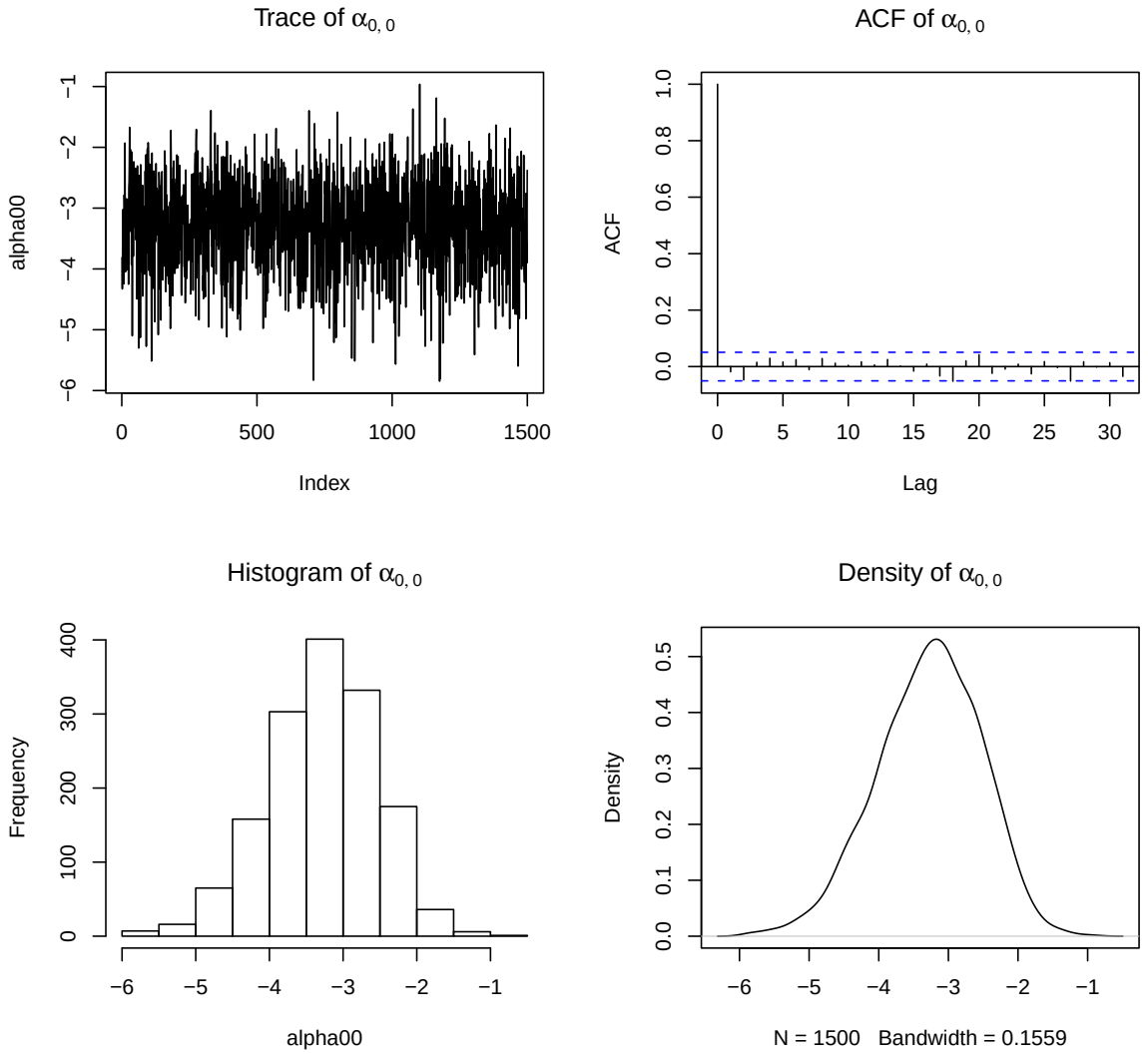

Figure C.21: Trace, autocorrelation function, histogram and density of $\alpha_{0,0}$ parameter of model 1

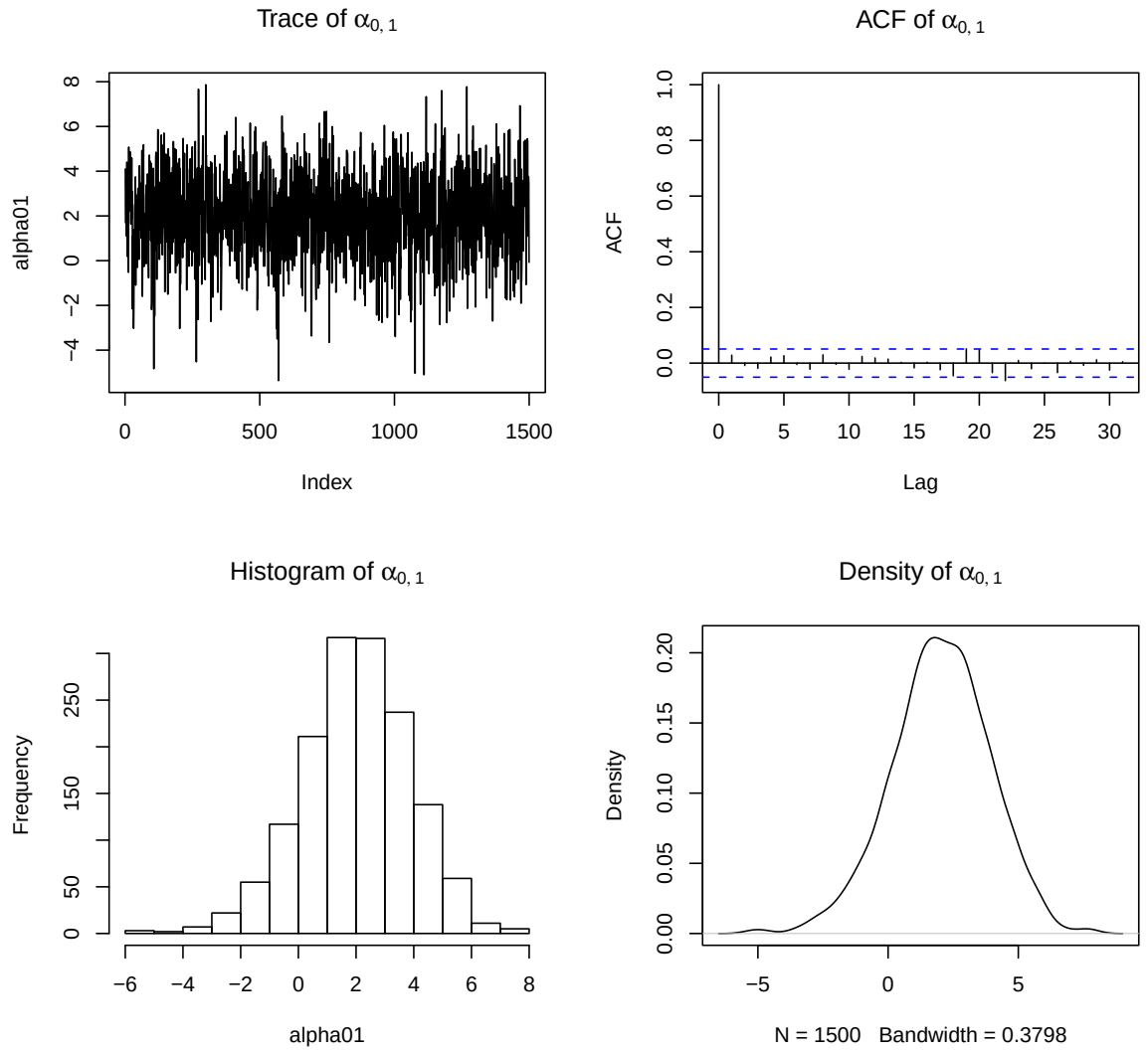

Figure C.22: Trace, autocorrelation function, histogram and density of $\alpha_{0,1}$ parameter of model 1

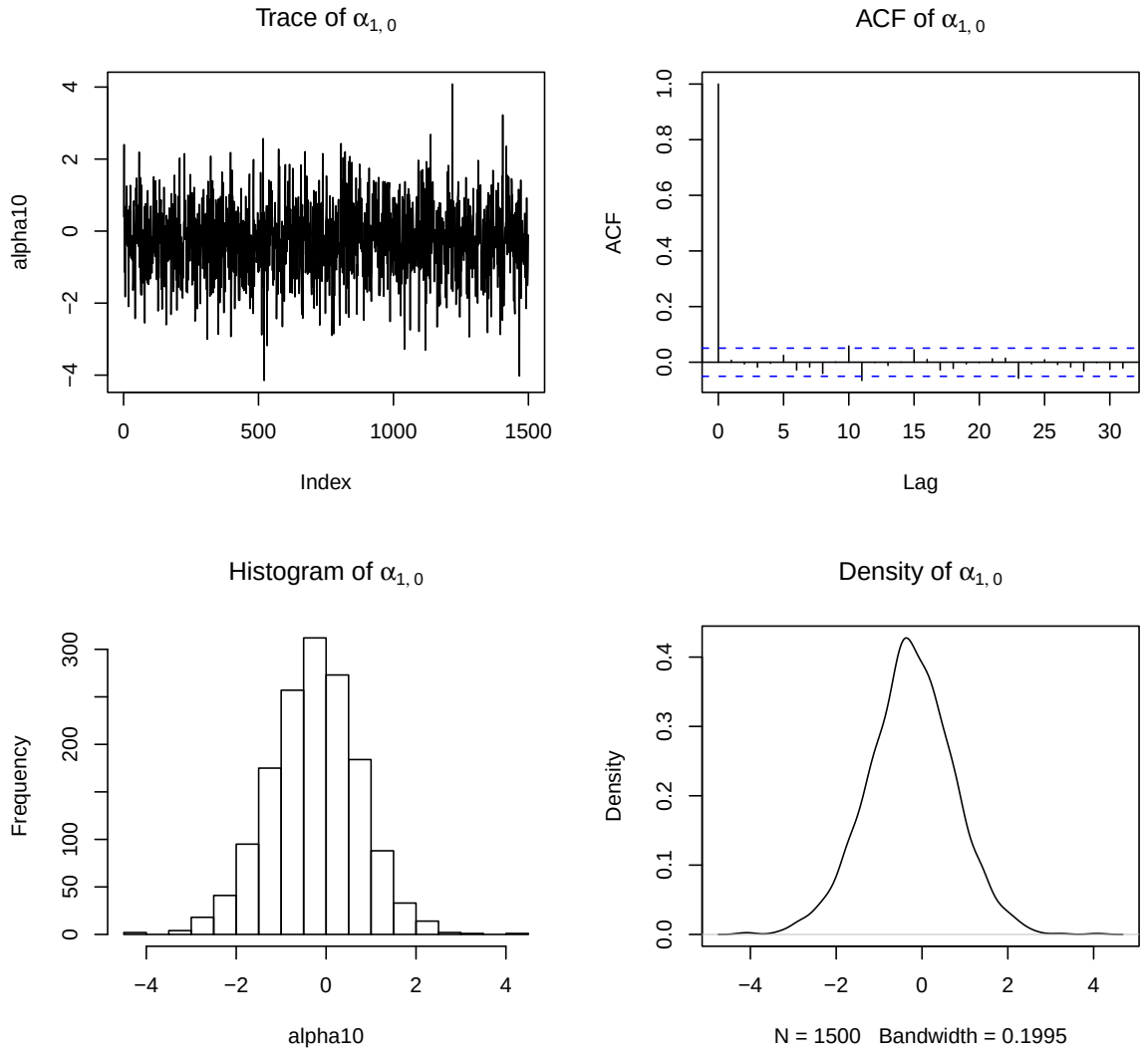

Figure C.23: Trace, autocorrelation function, histogram and density of $\alpha_{1,0}$ parameter of model 1

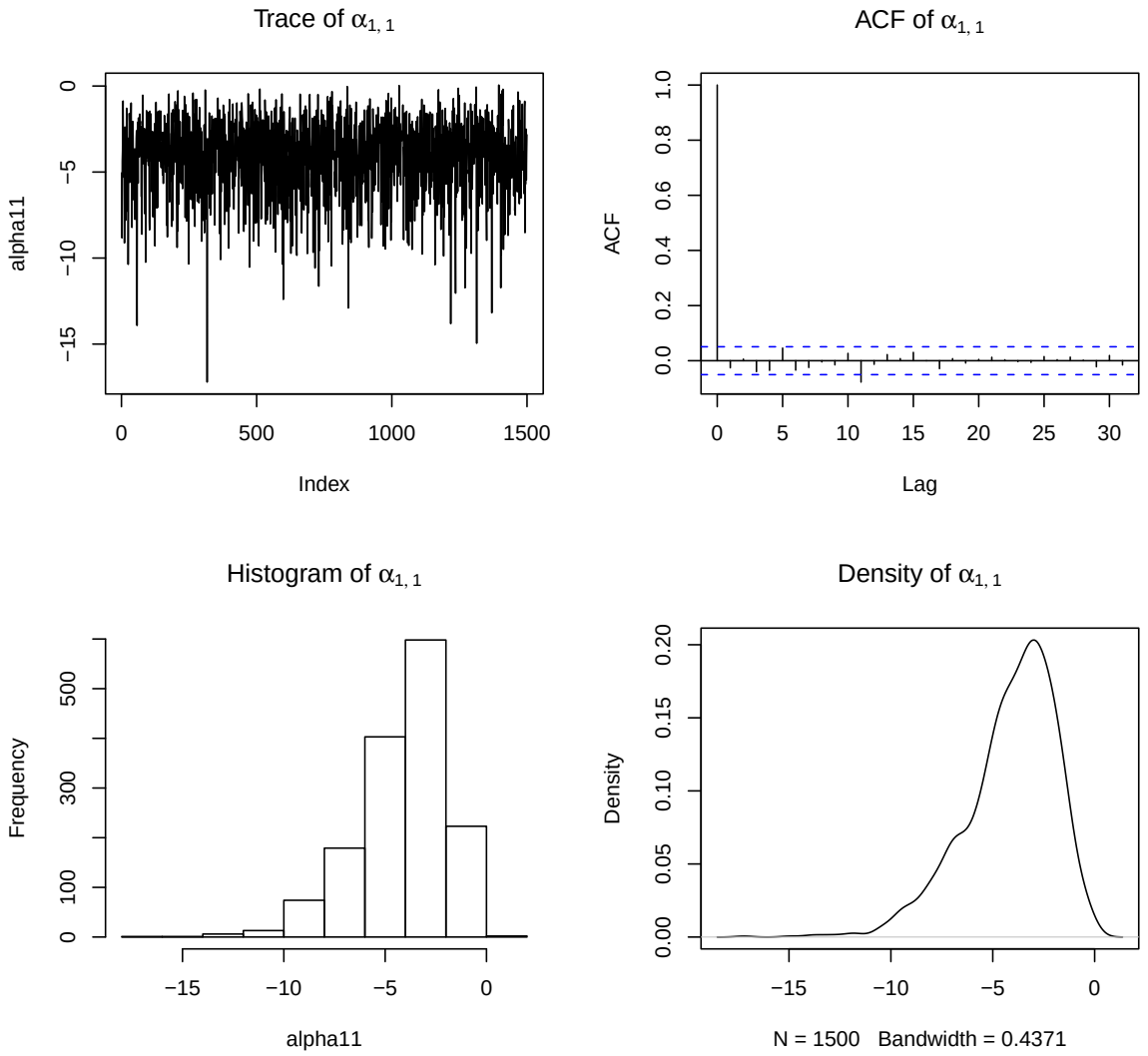

Figure C.24: Trace, autocorrelation function, histogram and density of $\alpha_{1,1}$ parameter of model 1

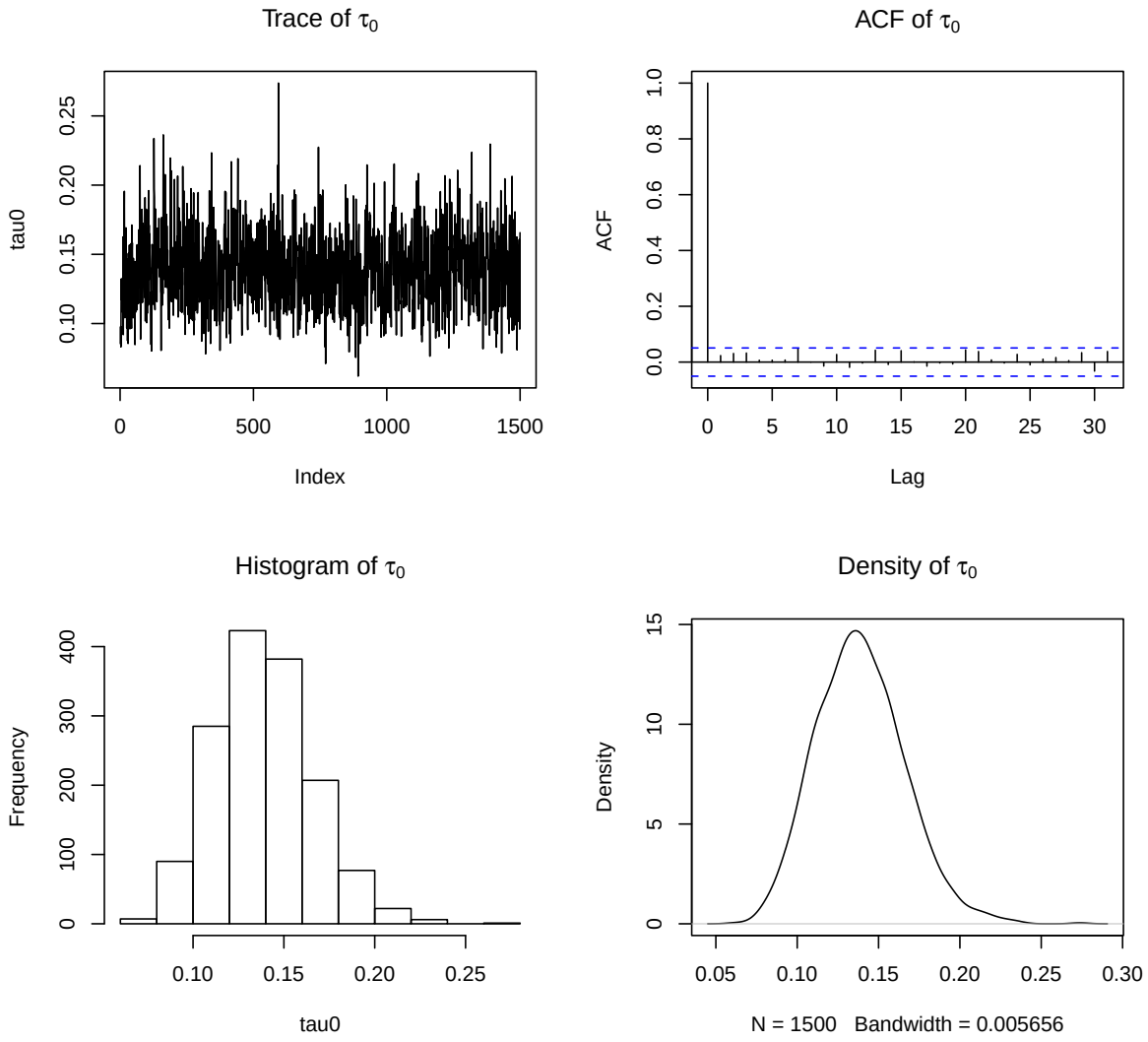

Figure C.25: Trace, autocorrelation function, histogram and density of τ_0 parameter of model 2

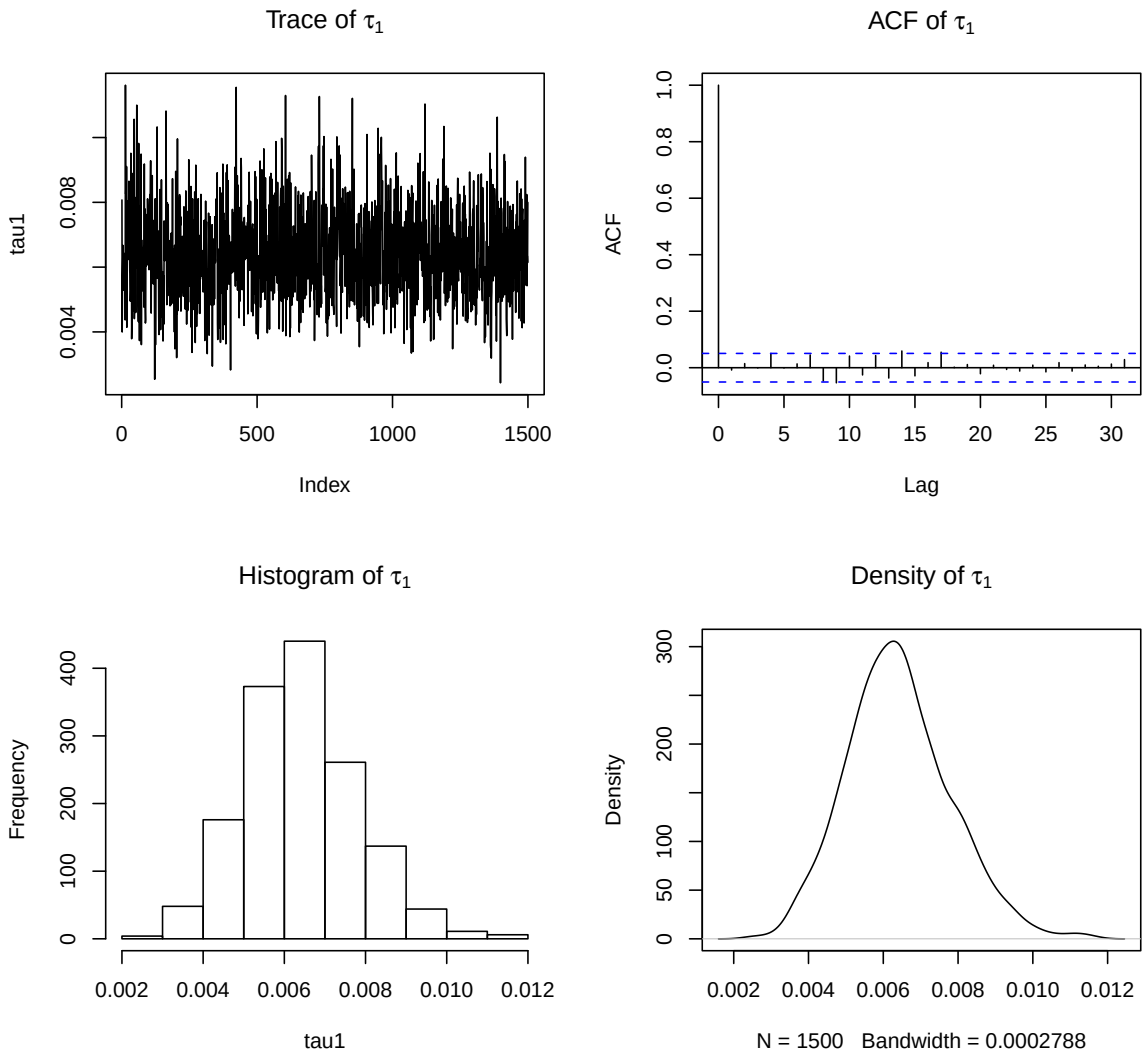

Figure C.26: Trace, autocorrelation function, histogram and density of τ_1 parameter of model 2

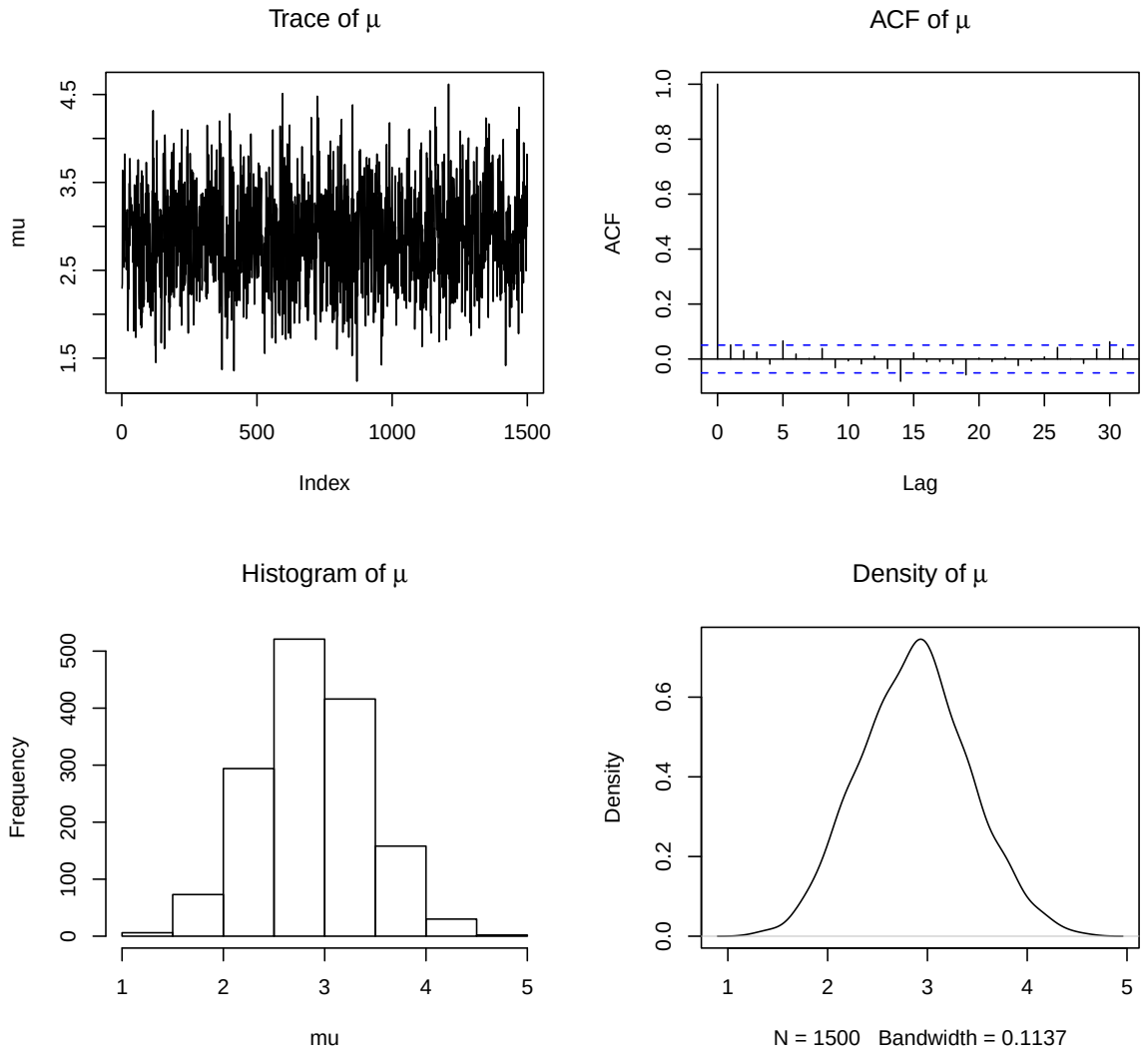

Figure C.27: Trace, autocorrelation function, histogram and density of μ parameter of model 2

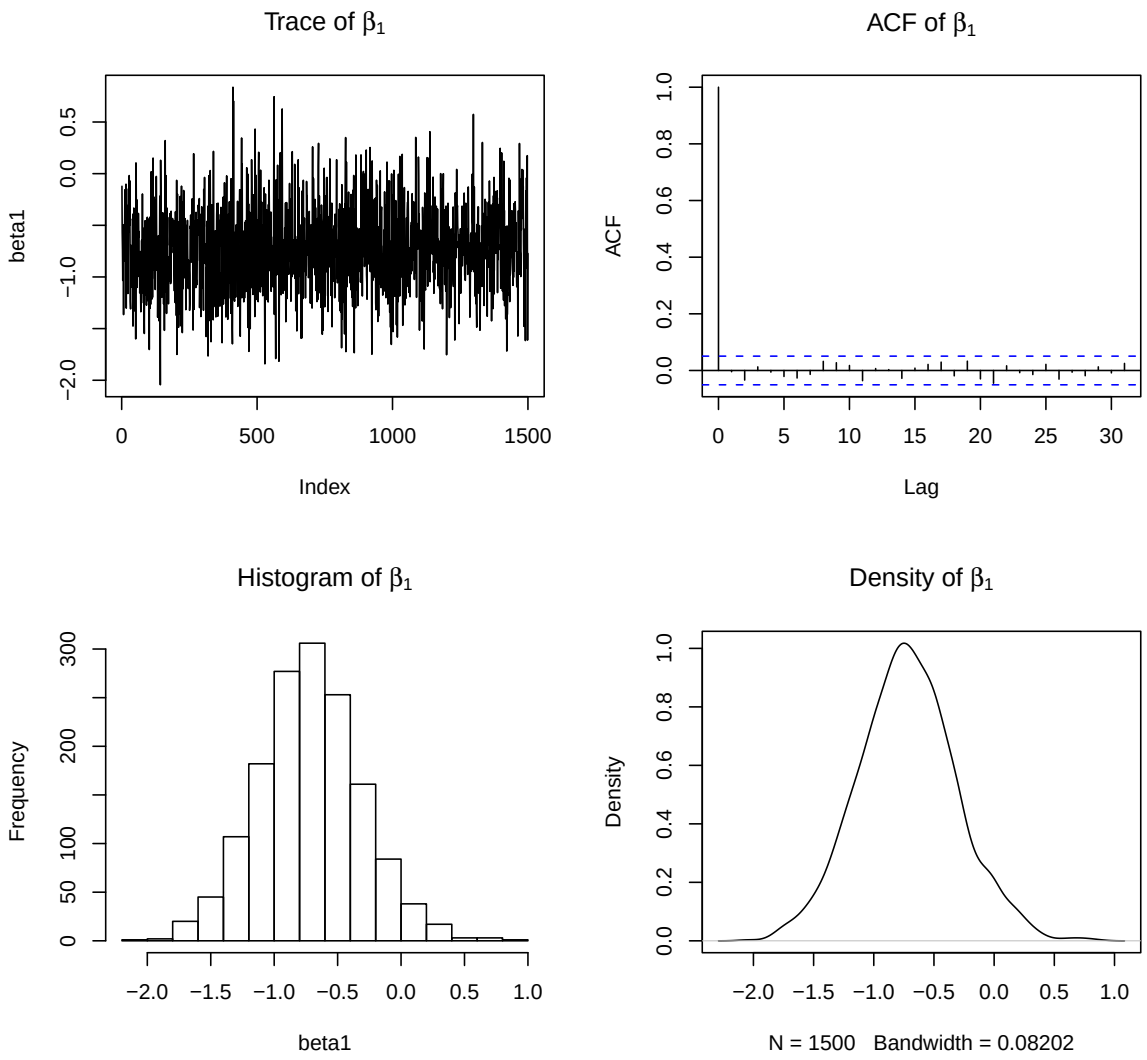

Figure C.28: Trace, autocorrelation function, histogram and density of β_1 parameter of model 2

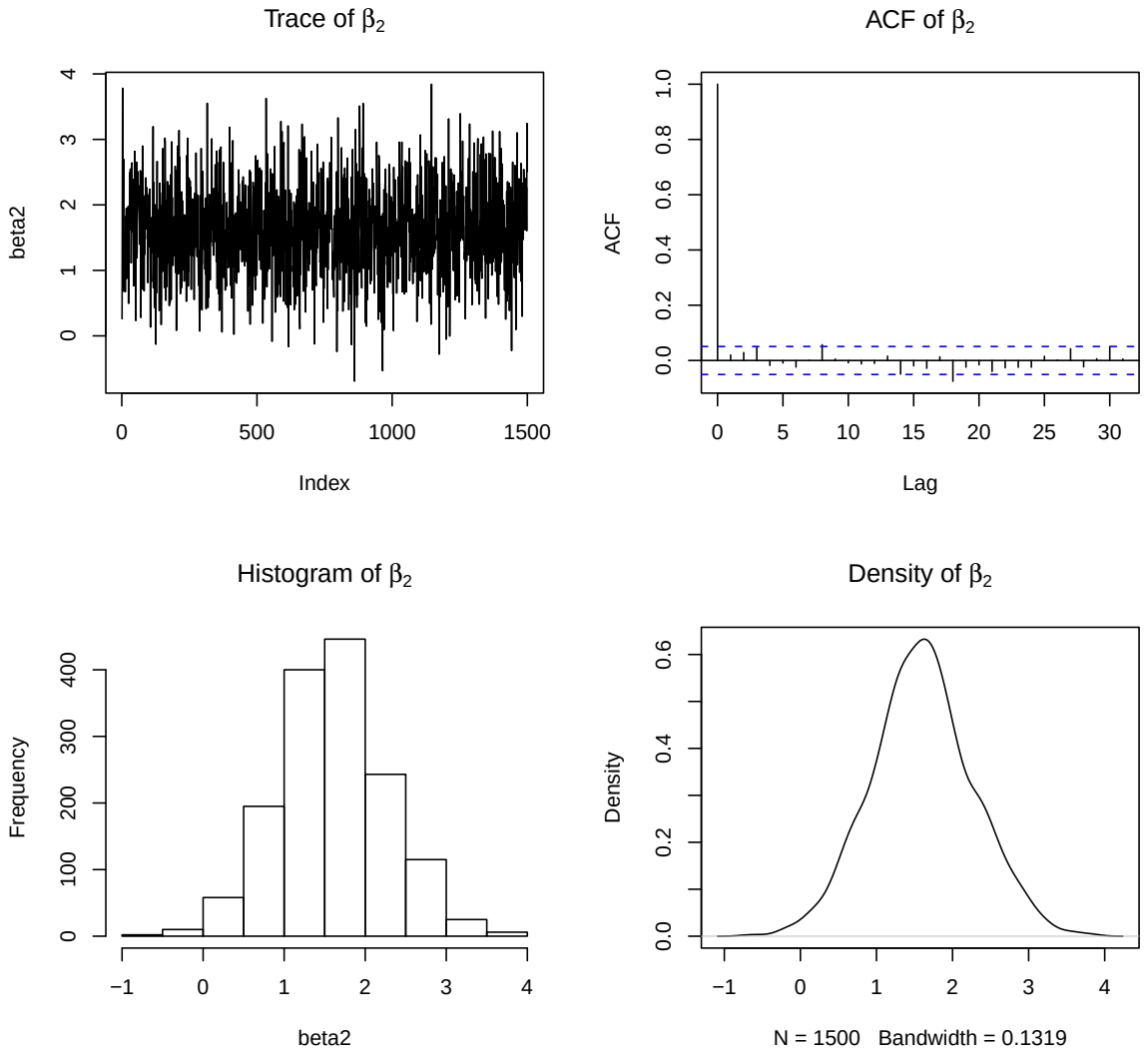

Figure C.29: Trace, autocorrelation function, histogram and density of β_2 parameter of model 2

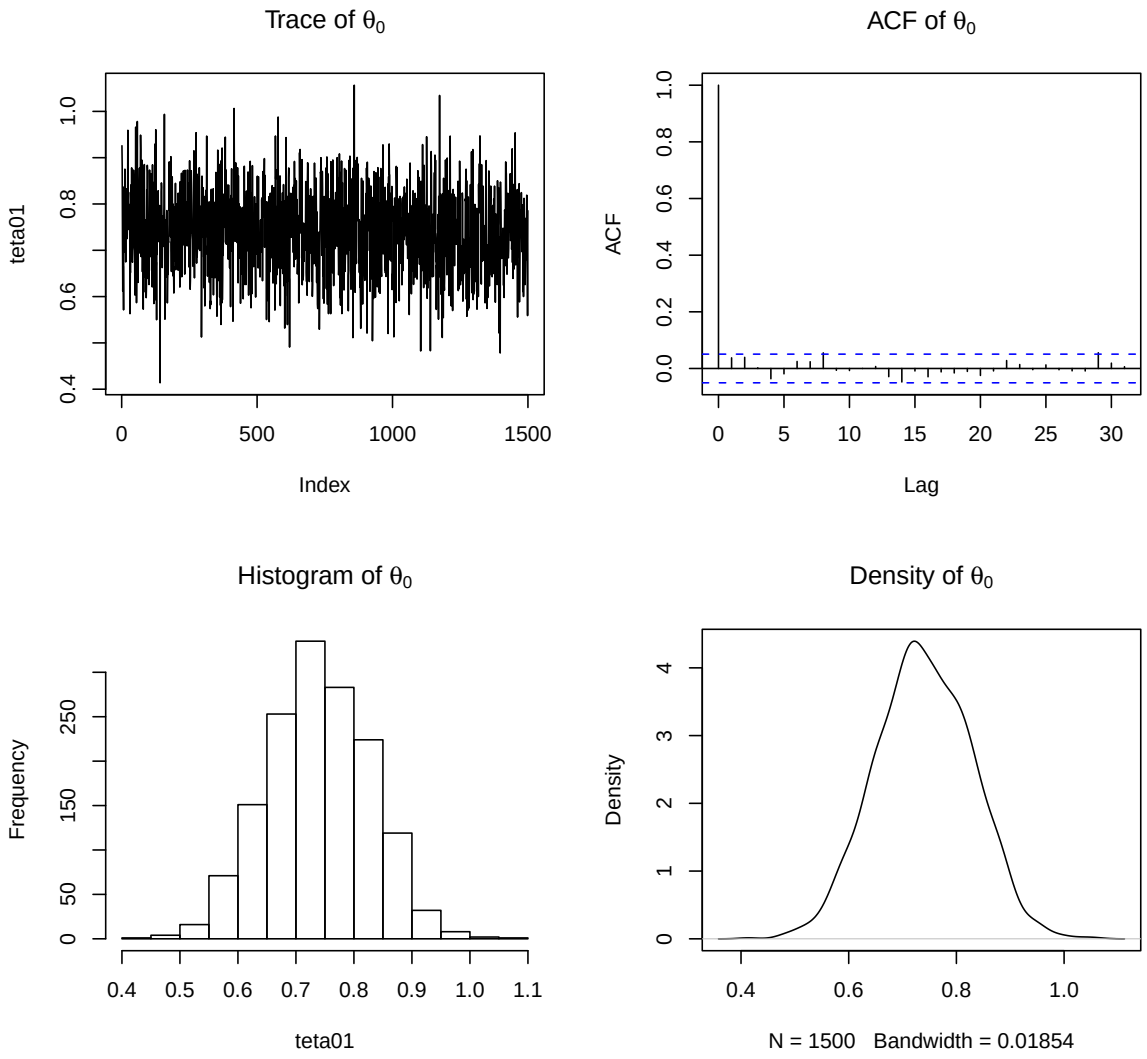

Figure C.30: Trace, autocorrelation function, histogram and density of θ_0 parameter of model 2

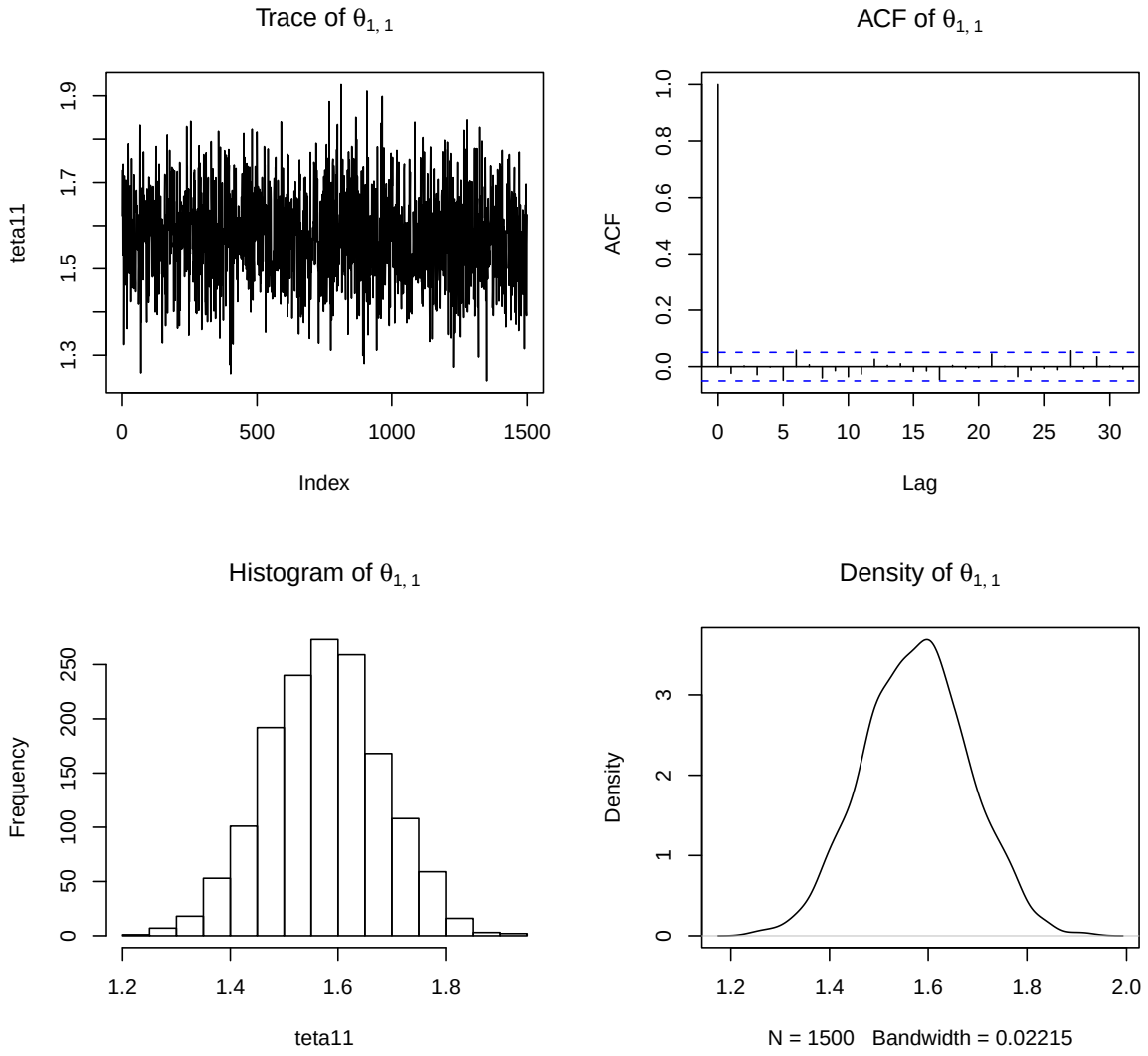

Figure C.31: Trace, autocorrelation function, histogram and density of $\theta_{1,1}$ parameter of model 2

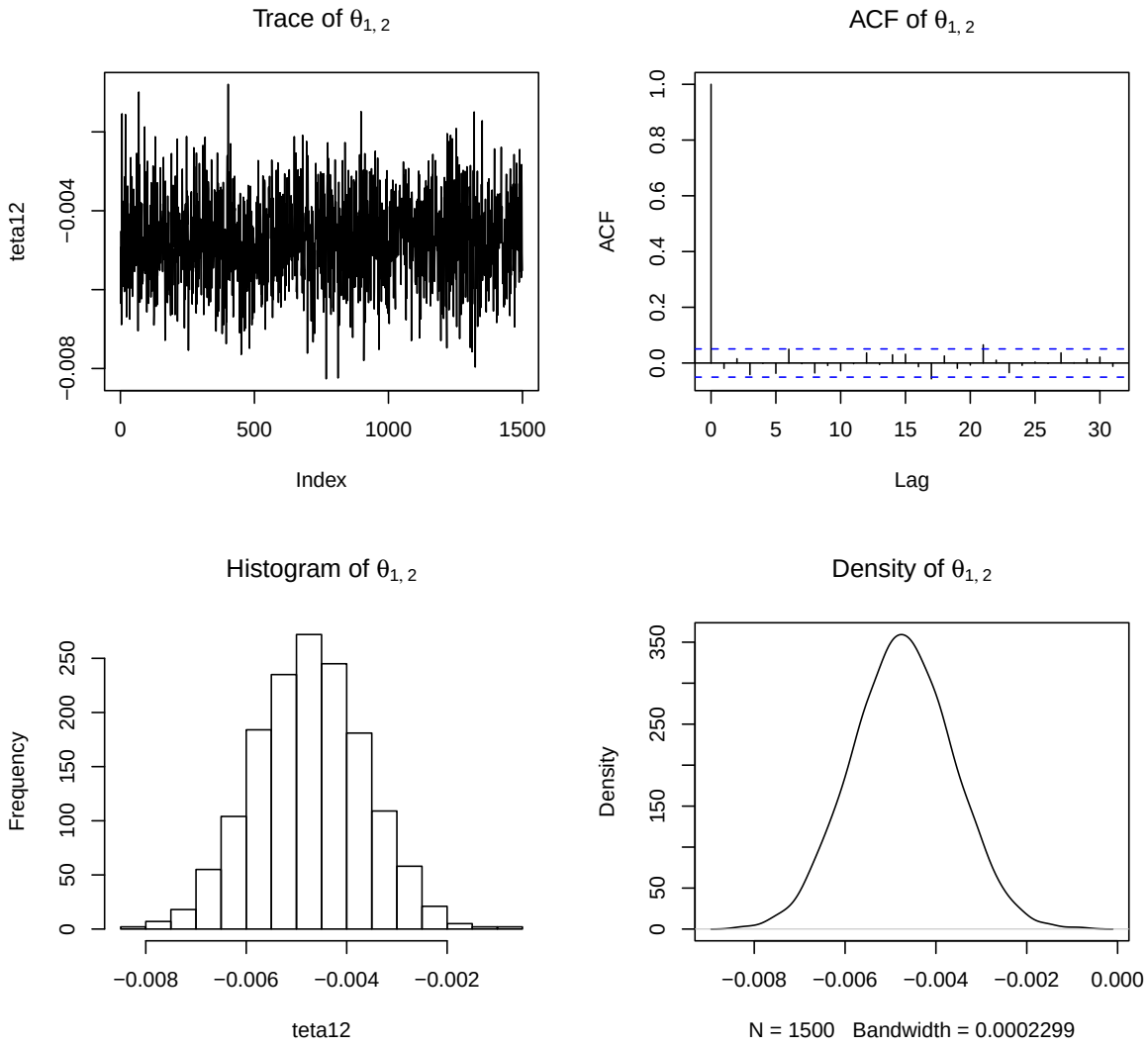

Figure C.32: Trace, autocorrelation function, histogram and density of $\theta_{1,2}$ parameter of model 2

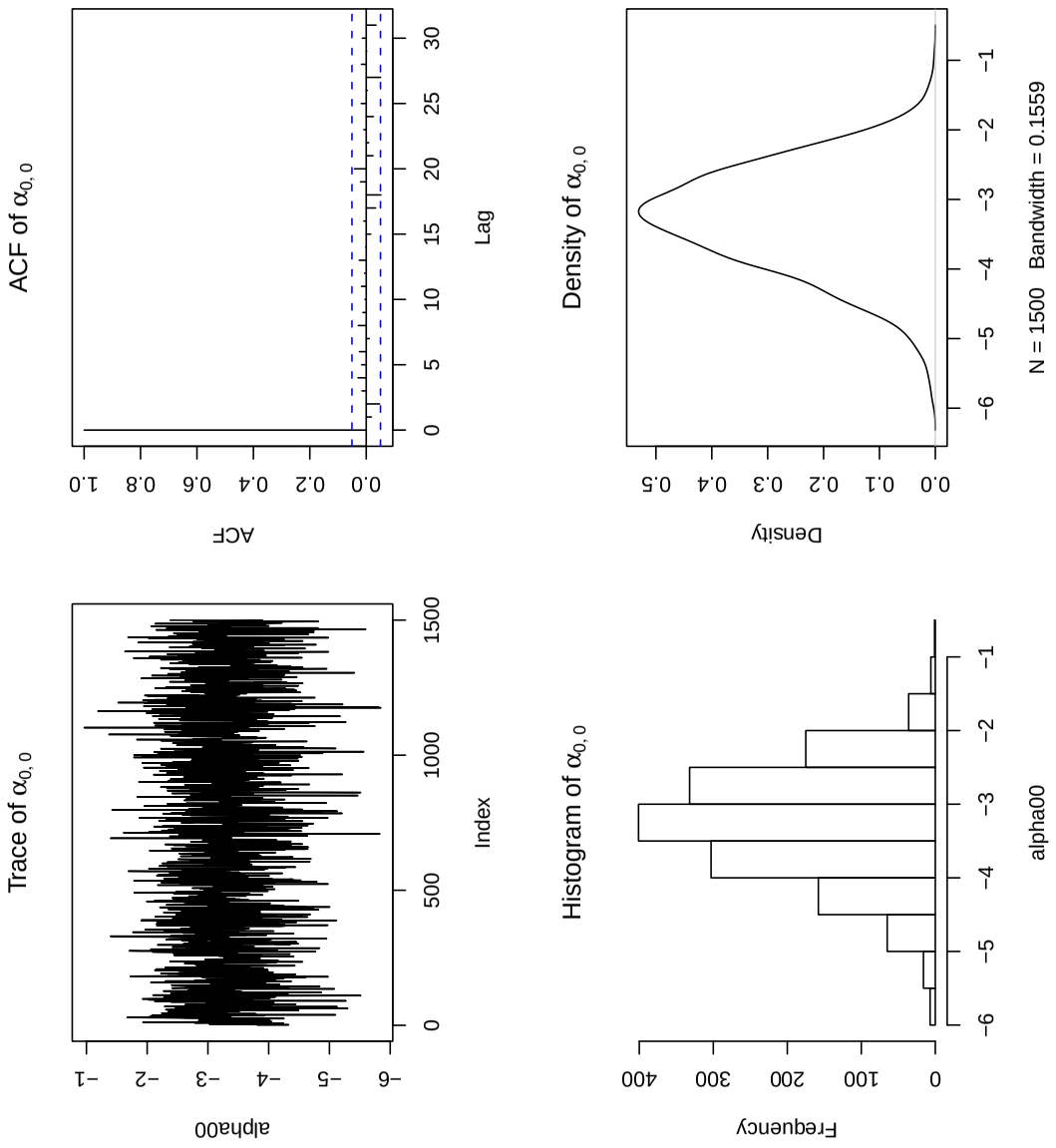

Figure C.33: Trace, autocorrelation function, histogram and density of $\alpha_{0,0}$ parameter of model 2

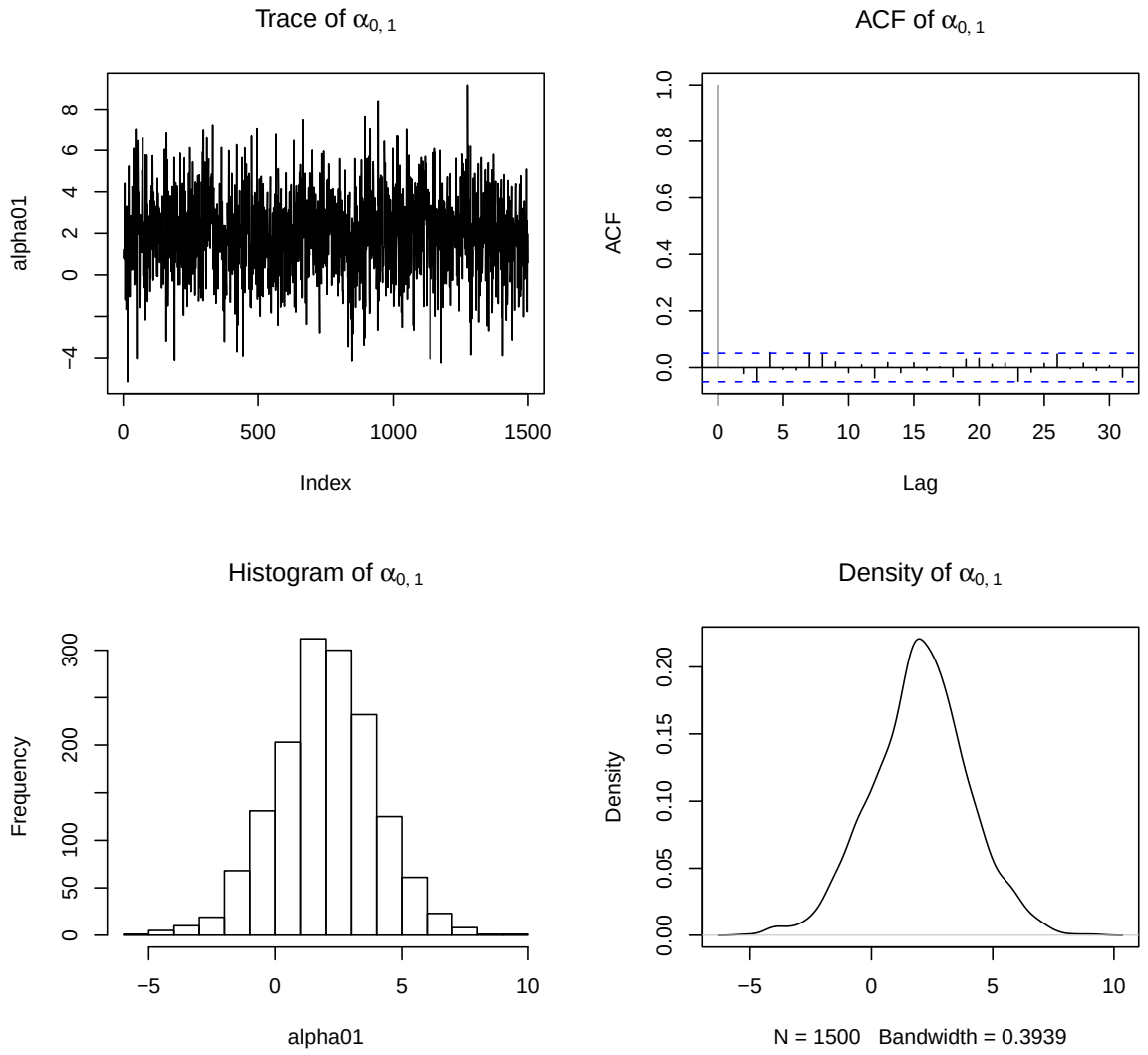

Figure C.34: Trace, autocorrelation function, histogram and density of $\alpha_{0,1}$ parameter of model 2

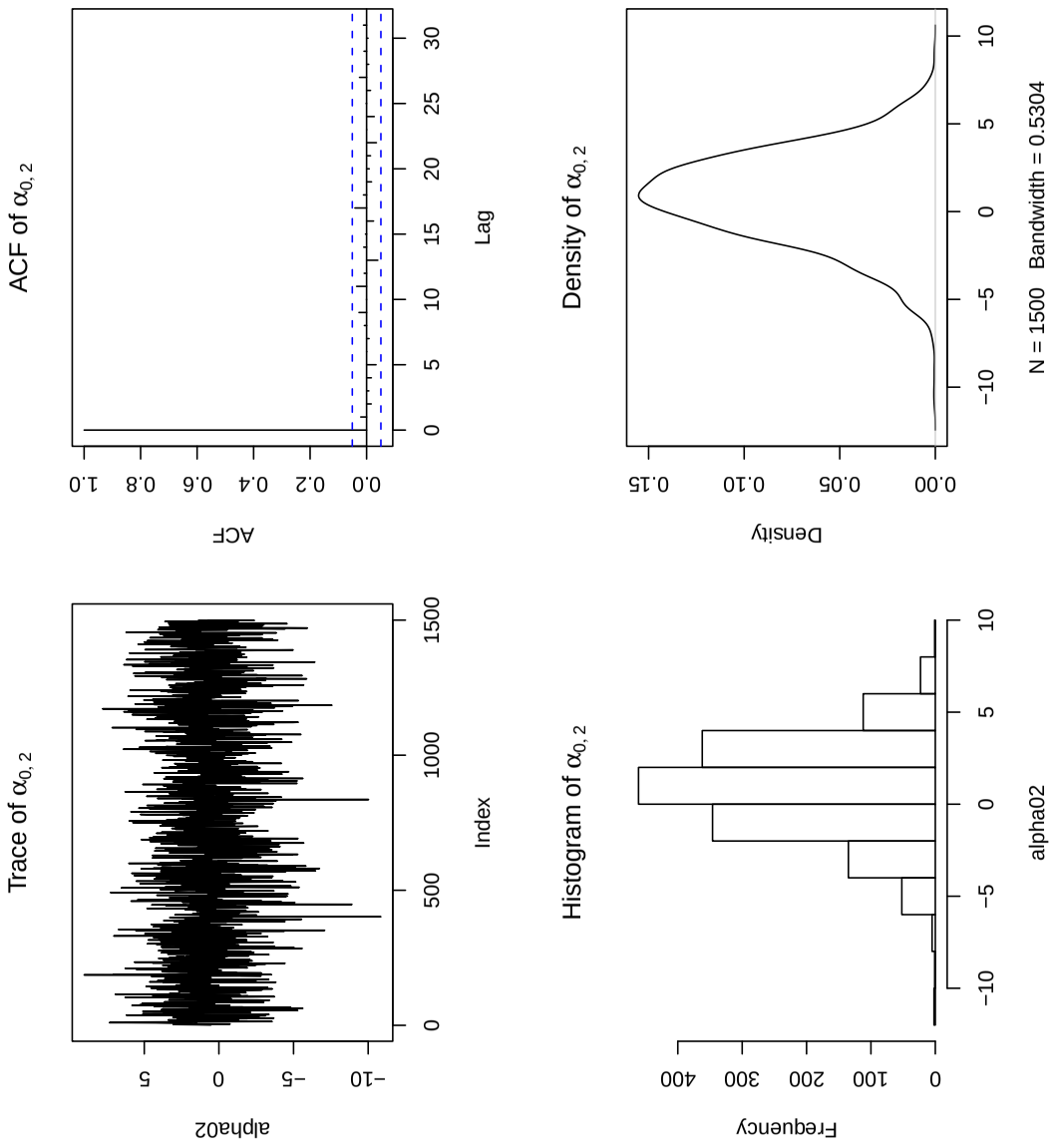

Figure C.35: Trace, autocorrelation function, histogram and density of $\alpha_{0,2}$ parameter of model 2

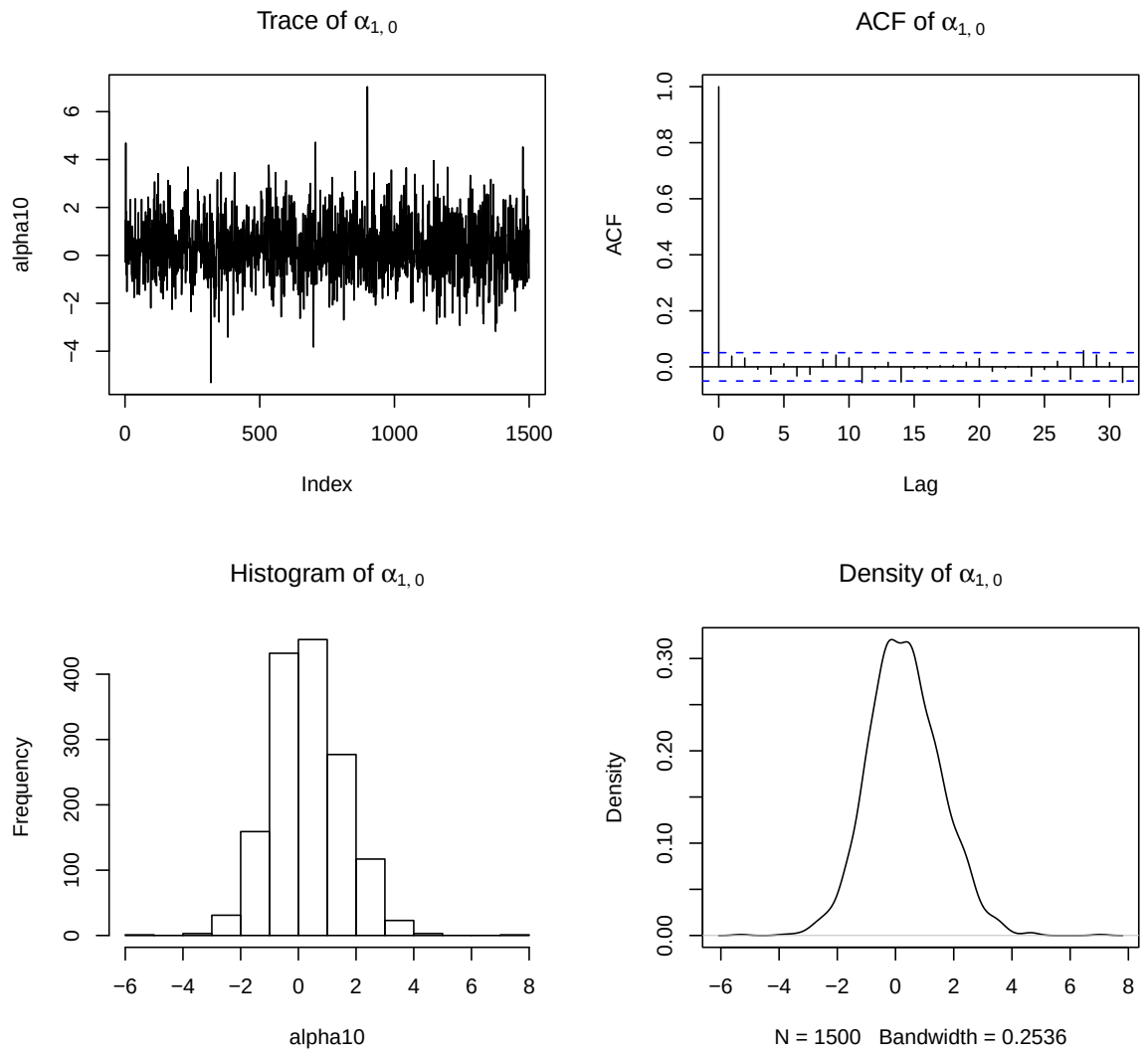

Figure C.36: Trace, autocorrelation function, histogram and density of $\alpha_{1,0}$ parameter of model 2

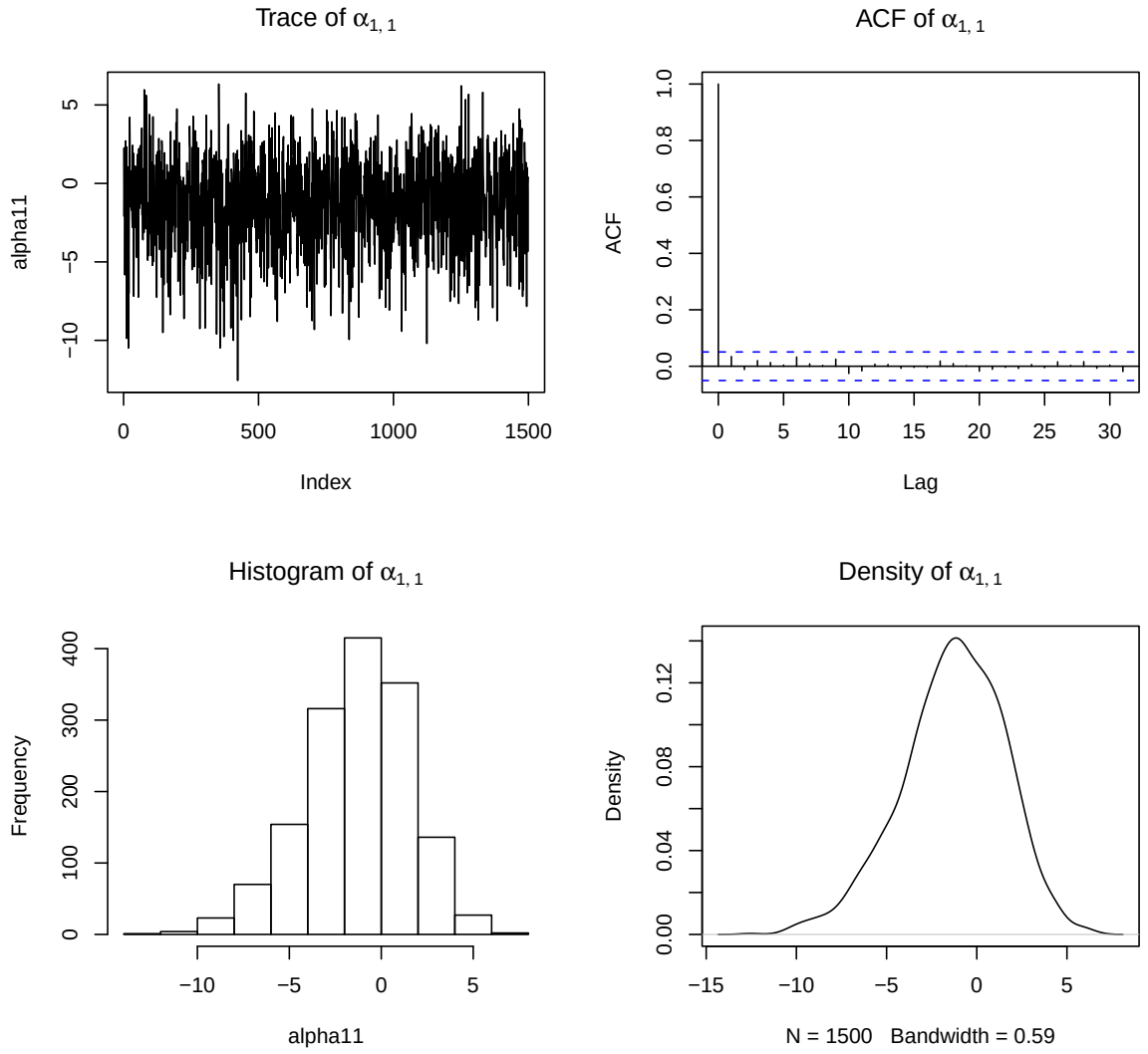

Figure C.37: Trace, autocorrelation function, histogram and density of $\alpha_{1,1}$ parameter of model 2

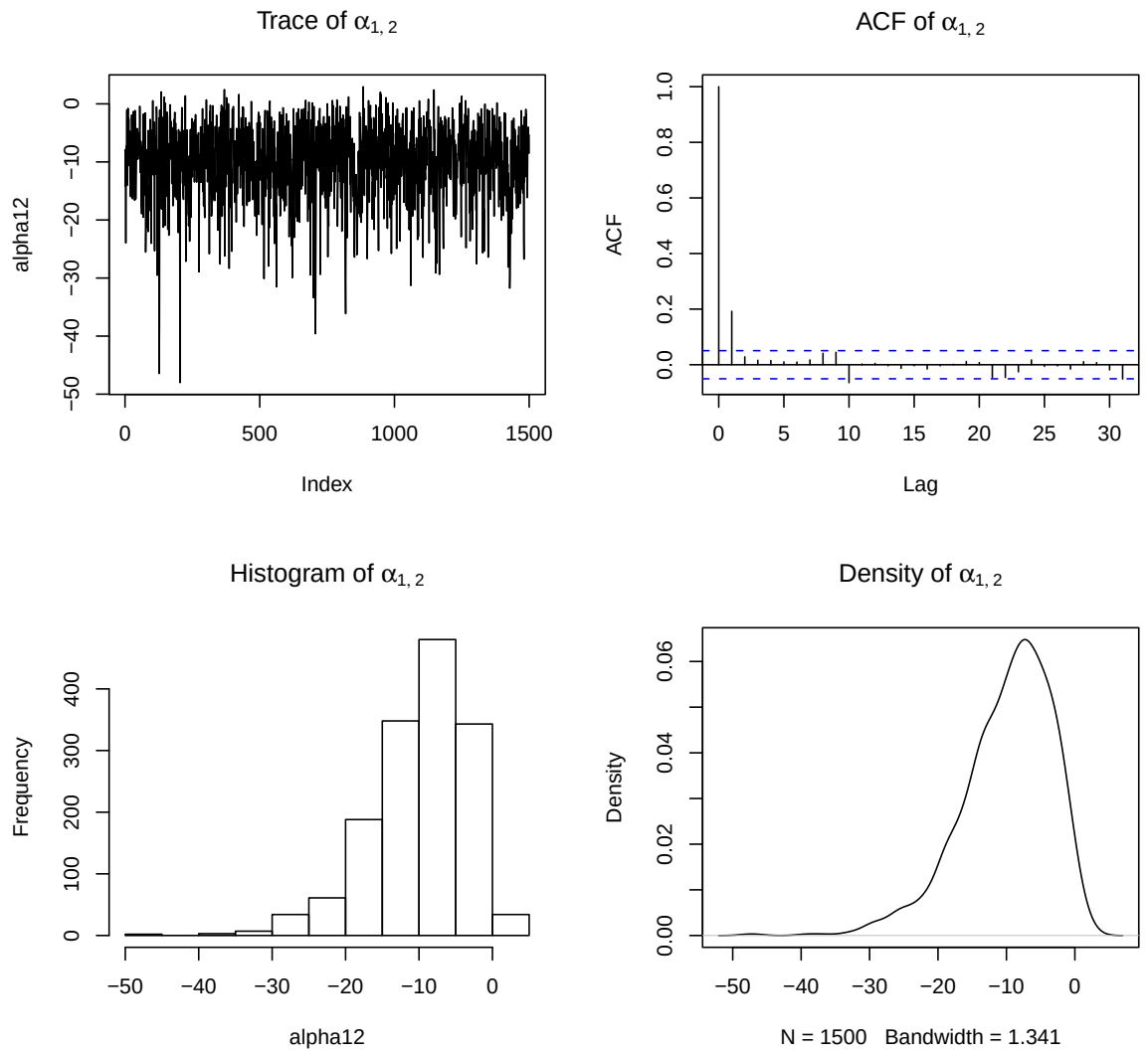

Figure C.38: Trace, autocorrelation function, histogram and density of $\alpha_{1,2}$ parameter of model 2